Verification of TIGGE Multi-model and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous US

Thomas M. Hamill

NOAA Earth System Research Laboratory, Physical Sciences Division

Boulder, Colorado USA

Submitted as an article

to

Monthly Weather Review

Revised, 5 December 2011

Corresponding author address:

Dr. Thomas M. Hamill
NOAA ESRL, Physical Sciences Division
R/PSD 1
325 Broadway
Boulder, CO 80305-3328
tom.hamill@noaa.gov
phone: (303) 497-3060 fax: (303) 497-6449

ABSTRACT

Probabilistic quantitative precipitation forecasts (PQPFs) were generated during July to October 2010 using European Centre (ECMWF), United Kingdom (UKMO), US (NCEP) and Canadian (CMC) forecast data. 24-hour accumulated precipitation forecasts were evaluated within the contiguous US against precipitation analyses for +1 to +5 days lead at 1-degree grid spacing.

PQPFs from ECMWF's ensembles generally had the highest skill of the raw ensemble forecasts, followed by CMC. PQPFs from CMC forecasts were the most reliable but the least sharp. PQPFs from NCEP and UKMO ensembles were the least reliable but sharper.

Multi-model PQPFs were more reliable and skillful than individual ensemble prediction system forecasts. The improvement was larger for heavier precipitation events than light events.

ECMWF ensembles were statistically post-processed using extended logistic regression and the five-member weekly reforecasts for the June - November period of 2002-2009. Multi-model ensembles were also post-processed using logistic regression and the last 30 days of prior forecasts and analyses. The reforecast-calibrated ECMWF PQPFs were more skillful and reliable for the heavier precipitation events than ECMWF raw forecasts but less sharp. Raw multi-model PQPFs were generally more skillful than reforecast-calibrated ECMWF PQPFs for the light precipitation events but about the same skill for the heavier events; also, they were sharper but somewhat less reliable than ECMWF reforecast-based PQPFs. Post-processed multi-model PQPFs did not provide as much improvement to the

raw multi-model PQPF as the reforecast-based processing did to the ECMWF forecast.

The evidence presented here suggests that all operational centers, even ECMWF, would benefit from generating reforecasts and sharing data in real time.

1. **Introduction**

An ongoing challenge with short- and medium range ensemble prediction systems (EPSs) is how to generate probabilistic forecasts that account for the system errors in the ensemble. System errors include sampling error due to the finite ensemble size, the error introduced by model imperfections such as the grid truncation, the use of deterministic parameterizations (Houtekamer and Mitchell 2005), and assimilation system and observation imperfections. There are many methods for treating system error, from introducing stochastic aspects into the ensemble prediction system (Buizza et al. 1999, Shutts 2005, Berner et al. 2009, Palmer et al. 2009, Charron et al. 2010), using multiple parameterizations (Charron et al. 2010, Berner et al. 2011), using multiple models (Bougeault et al. 2010), and statistical post-processing.

Two methods that will be explored and contrasted here are the multi-model methods and statistical post-processing. The underlying hypothesis of multi-model ensembles (Krishnamurti et al. 2000, Wandishin et al. 2001, Mylne et al. 2002, Doblas-Reyes et al. 2005, Hagedorn et al. 2005, Weigel et al. 2008, Candille 2009, Johnson and Swinbank 2009, Bougeault et al. 2010, Iversen et al. 2011) is that the many differences between constituent EPSs will result in them generating ensemble forecasts with quasi-independent systematic errors, so the combination may result in a more accurate estimate of the uncertainty. Practically, also, these are ensembles of opportunity. If all centers are willing to share rather than sell their forecast data, the additional members can be used for only the cost of data transmittal and storage, so they may provide an inexpensive way to improve forecast skill. However, there are some potential disadvantages of multi-model ensembles. Developing an accurate, stable weather prediction system is costly, so multi-model ensembles are likely to be less useful when formed from immature systems. System outages may prevent routine access to other centers' ensembles. One or other of the models is likely to have been changed recently, rendering it difficult to understand the multi-model system error characteristics. Also, the hypothesis of quasi-independent errors may not always hold. Practically, each operational center is interested in providing a high-quality product without depending on another center's data. When another center develops a method that improves the forecast significantly, it may be adopted at other operational centers. The similarity could result in some co-linearity of errors and decreased collective usefulness (Lorenz et al. 2011).

Another method for addressing system error is through statistical post-processing. Discrepancies between time series of past forecasts from a fixed model and the verifying observations/analyses can be used to modify the real-time forecasts. For some variables such as short-range forecasts of surface temperature, a short time series may be sufficient (Stensrud and Yussouf 2003, Yussouf and Stensrud 2007, Hagedorn et al. 2008). For others such as heavy precipitation and longer-lead forecasts, using a long time series of reforecasts has been shown to dramatically improve the reliability and skill of the probabilistic forecasts (Hamill et al. 2004, Hamill et al. 2006, Hamill and Whitaker 2007, Wilks and Hamill 2007, Hamill et al. 2008). A drawback of using reforecasts is that a forecast time series spanning many years or even decades may be necessary to produce a sufficiently large sample to adjust for systematic errors in rare-event forecasts. Since forecast models are frequently updated, which may change the systematic error characteristics, either a forecast model must be frozen once a reforecast data set has been generated, or a new reforecast data set must be generated every time the modeling system changes significantly. Hence, reforecasting can be computationally expensive and can restrict the ability of a forecast center to upgrade its system rapidly. Recently, statistical post-processing methods have been the subject of much investigation (Gneiting et al. 2005, Raftery et al. 2005, Sloughter et al. 2007, Wilson et al. 2007, Vannitsem and Nicolis 2008, Glahn et al. 2009, Bao et al. 2010).

To date, however, there have been no systematic comparisons of multi-model and reforecast-calibrated PQPFs verified over a large enough area and a long enough period of time to confidently assess the relative strengths and weaknesses of these two approaches. This study attempts to provide such a comparison for this high-impact forecast parameter. Using TIGGE forecast data from the US National Centers for Environmental Prediction (NCEP), the Canadian Meteorological Centre (CMC), the United Kingdom Met Office (UKMO), and the European Centre for Medium-Range Weather Forecasts (ECMWF), multi-model ensemble 24-h accumulated probabilistic forecasts of precipitation were generated and then compared against ECMWF forecasts that were statistically adjusted using their reforecast data set. The comparison was performed over the contiguous US (CONUS) during the period July-October 2010. Statistical adjustments were also attempted with multi-model forecasts, trained on the previous 30 days of forecasts and analyses.

Below, section 2 describes the data sets used in this experiment, the verification methodology, and the statistical post-processing method. Section 3 provides results, and section 4 some conclusions.

2. **Data sets and methods**.

a. Analysis data used.

A recently created precipitation data set, NCEP's Climatology-Calibrated Precipitation Analysis (CCPA), was used for verification. The CCPA attempts to combine the relative advantages of the 4-km, hourly NCEP Stage-IV precipitation analysis (Lin and Mitchell 2005), and the daily, 0.25-degree NCEP Climate Prediction Center (CPC) Unified Precipitation Analysis (Higgins et al. 1996). The former is based on gauge and radar data, the latter solely on gauge data. A disadvantage of the Stage-IV product is that it may inherit some of the biases due to the estimation of rainfall from radars. A disadvantage of the CPC product is that there are areas of the US that are only sparsely covered by gauge data. The CCPA analysis regressed the Stage-IV analysis (the predictor) to the CPC analysis (the predictand), thereby reducing bias with respect to the in-situ observations. In several of the driest locations in the western US, the CCPA analysis was set to missing, for the regression analysis was untrustworthy and singular due to no precipitation in either analysis product. In such cases, the CCPA analysis for this study was simply replaced with the Stage-IV analysis. For our purposes, we used CCPA analyses that also were upscaled to 1 degree and accumulated over a 24-h period in a manner that preserved total precipitation, similar to the “remapping” procedure described in Accadia et al. (2003). The CCPA analyses were available from 2002 – current, a shorter period than the ECMWF reforecasts, thus limiting the amount of training data that could be used in the statistical post-processing.

b. Forecast and reforecast model data.

For this experiment, 20 perturbed member forecasts of 24-h accumulated precipitation were extracted from the UKMO, CMC, NCEP, and ECMWF ensemble systems archived in the TIGGE database at ECMWF. Probabilities were calculated directly from the ensemble relative frequency, referred to as “raw” probabilities henceforth. The forecast period was July to October 2010; only 00 UTC initial time forecasts were extracted in order to allow comparison with a post-processed forecasts using ECMWF’s reforecasts, which were generated only from 00 UTC initial conditions. Daily forecasts of 24-h accumulated precipitation were examined from +1 to +5 day lead. Regardless of the original model resolution, all centers’ forecasts were bi-linearly interpolated to a 1-degree latitude-longitude grid covering the CONUS using ECMWF’s TIGGE portal software. ECMWF’s interpolation procedure set the amount to zero if there was no precipitation at the nearest neighboring point and the interpolated value was less than 0.05 mm. No control forecasts were included, just the forecasts from the perturbed initial conditions. Other forecast centers’ contributions to the TIGGE archive were not used here for various reasons, such as the unavailability of 00 UTC ensemble forecasts from the Japan Meteorological Agency. For size consistency and to facilitate skill comparisons, only the first 20 of the full 50 ECMWF member forecasts were used in the generation of the multi-model ensemble, though the 50-member ECMWF forecasts were evaluated for skill and reliability.   More detailed descriptions of the configuration of these four ensemble systems are described in Appendix 1.

When calibrating ECMWF data with reforecasts, the 5-member weekly reforecasts precipitation data were extracted from ECMWF's weekly reforecast archive (Hagedorn 2008) and similarly interpolated to the 1-degree grid.   The control reforecasts were initialized from the ERA-Interim reanalysis (Dee et al. 2011), which used version Cy31r2 of the ECMWF Integrated Forecast System (IFS) in the data assimilation process. The 2010 real-time ensemble forecasts and the reforecasts were then run using IFS model version Cy36r2 (more detail is provided in appendix A).

The four perturbed initial conditions for the reforecasts were generated with a combination of their singular-vector approach (Molteni et al. 1996, Barkmeijer et al. 1999) and their "ensembles of data assimilations" or "EDA" (Isaksen 2010) that used a cycled, reduced resolution 4D-Var and perturbed observations. However, for initialization of the reforecasts, ECMWF used the EDA perturbations from 2010 and applied them to the 2002-2009 data rather than running the EDA during the reforecast period. To apply EDA to dates in the past would have been computationally expensive, but having not done so may have resulted in the perturbations in the reforecast having less flow-dependent character, possibly making them somewhat statistically inconsistent with ECMWF's real-time ensemble forecasts.

Since precipitation analysis data was only available for the period from 2002 forward, the training data was limited to the reforecasts for period of June to November, 2002-2009, or 8 years. To limit the possible deleterious effects of seasonal biases in the forecast model, only the reforecast data for the month of interest and the month before and after were used. For example, when calibrating July forecasts, June-July-August reforecast data was used. With reforecasts generated once per week, this typically meant there were ~13 once-weekly samples × 8 years = 104 samples. Tthis was in many cases an insufficient sample size, so data from other grid points were used to increase the training sample size (appendix B).

c. Statistical post-processing methodology.

The extended logistic regression (ELR) approach of Wilks (2009) was used here, a procedure that permitted the development of a single regression equation that was suitable for predicting probabilities of exceeding any precipitation amount. The probability was estimated with a function of the form

$$p = \frac{\exp\left[f\left(\mathbf{x}\right)\right]}{1+\exp\left[f\left(\mathbf{x}\right)\right]}, \qquad (1)$$

where $f(\mathbf{x})$ was a linear function of the predictor variables. In this case, the predictors were (a) the ensemble-mean forecast $\overline{x}$ raised to the 0.4 power, (b) the product of (a) and the variance σ^2 to the 0.4 power, and (c) the precipitation event threshold T raised to the 0.4 power. The linear function was thus

$$f(\mathbf{x}) = b_0 + b_1\overline{x}^{0.4} + b_2\overline{x}^{0.4}\sigma^{2\times 0.4} + b_3T^{0.4}\,. \qquad (2)$$

The choice of these predictors was arrived at through some trial and error. The power transformation of the predictors helped make the input data somewhat more normally distributed. The probabilistic forecast skill was also only mildly dependent on the inclusion/exclusion of the predictor with the product of the transformed mean and variance. Skill was also only slightly dependent on the power of the transform, with 0.4 providing an approximate minimum. Previous values of power transformations in the literature have ranged from ½ in Hamill and Whitaker (2006) and Schmeits and Kok (2010), 1/3 in Sloughter et al. (2007), and ¼ in Hamill et al. (2008) and Roulin and Vannitsem (2011). The use of the product of the ensemble mean and variance follows Wilks and Hamill (2007). The additional predictor incorporating *T* permitted the single regression equation to be used to predict probabilities across the range of possible amounts. A disadvantage of this ELR approach (as opposed to approaches such as the analog approach discussed in Hamill and Whitaker (2006)) was that this algorithm was not able to correct for possible position biases in forecast features.

ELR was applied both to calibrating real-time multi-model forecasts and to calibrating ECMWF forecasts alone using the weekly reforecasts. It was found that forecast skill increased if some method was applied to increase the modest training sample sizes. A discussion of how sample sizes were augmented using data from other nearby forecast grid points is provided in Appendix 2.

Roulin and Vannitsem (2011) noted that since the ECMWF reforecast size (5 members) was smaller than the operational ensemble size (50 members; or in the case here, 20 members selected from the 50), the regression coefficients may be somewhat biased when trained with a smaller ensemble compared to what they would be were they trained with a larger ensemble. Hence, when the coefficients are used to correct the larger real-time ensemble, they may produce somewhat biased probabilistic forecasts. They adjusted the values of the 5-member ensemble training data to better estimate the values that would be obtained with the larger real-time ensemble. An analogous approach was tried here but did not improve the forecast skill. The results discussed below will omit this adjustment.

d. Verification methods.

The primary verification methods used here were Brier Skill Scores (*BSS*), continuous ranked probability skill scores (*CRPSS*), and reliability diagrams (Wilks 2006). The *BSS* and *CRPSS* as conventionally calculated (see section 7.4.2 of Wilks (2006)) can exaggerate forecast skill, attributing skill to variations in climatological event probabilities. Thus, the procedures suggested in Hamill and Juras (2006) were used here to avoid this.

To calculate the *BSS*, the score was calculated separately for subsets of points that had more uniform climatological probabilities. The overall *BSS* was the average of the skill scores over these subsets. The specific procedure was as follows. Using the 1-degree precipitation analysis data from 2002-2009, for each month the climatological probability of a given precipitation event was estimated from the observed frequency. For a given event such as > 1 mm (24 h)$^{-1}$, the n_s grid points within the CONUS were sorted from lowest to highest event probability. The sorted points were then divided into k=6 classes, with the lowest bin containing the ~ $n_s/6$

grid points with the lowest event probabilities, the highest bin containing the n_s /6 points with the highest probabilities, and so on (Fig. 1). Let $\mathbf{BS}^{f1} = \left[\mathbf{bs}_1^{f1}, \ldots, \mathbf{bs}_6^{f1} \right]$ denote a matrix of Brier scores for forecast model *f1*, where $\mathbf{bs}_i^{f1}$ was a n_d – dimensional (= 123, the number of case days here) column vector of average Brier scores for the points in the *i*th class and for forecast model *f1*. An element of this vector thus provided the average Brier Score for all of the grid points in the *i*th class on a particular day; the samples were weighted by the cosine of their latitude to account for differences in grid box size. The average over the 123 case days produced a vector $\overline{\mathbf{bs}}^{f1} = \left[\overline{bs}_1^{f1}, \ldots, \overline{bs}_6^{f1} \right]$. Similarly, for climatology there was an array of Brier scores, $\mathbf{BS}^c = \left[\mathbf{bs}_1^c, \ldots, \mathbf{bs}_6^c \right]$ and a vector of their averages over the 123 days, $\overline{\mathbf{bs}}^c = \left[\overline{bs}_1^c, \ldots, \overline{bs}_6^c \right]$. Following Hamill and Juras (2006) eq. (9), the overall *BSS* for model *f1* was then calculated as

$$BSS = \sum_{k=1}^{6} \frac{1}{6} \left(1 - \frac{\overline{bs}_k^{f1}}{\overline{bs}_k^c} \right). \qquad (3)$$

The boundaries between the classes were calculated independently for each event, so it was possible that a given grid point may have been assigned to different classes when evaluating, say, the 1- and 10-mm *BSS*s.

BSS confidence intervals were estimated using the paired block bootstrap approach of Hamill (1999). The input data to the bootstrap approach consisted of arrays of $\mathbf{BS}^{f1}$ and $\mathbf{BS}^{f2}$ for two competing models, *f1* and *f2*, as well as $\mathbf{BS}^c$. Let $\mathbf{bs}^{f1}(d) = \left[bs_1^{f1}(d), \ldots, bs_6^{f1}(d) \right]$ be the vector of forecast scores on the *d*th case day, and similarly $\mathbf{bs}^{f2}(d)$ the vector for forecast model *f2*. The daily differences in Brier scores, $\mathbf{bs}^{f1}(d) - \mathbf{bs}^{f2}(d)$ were determined to be approximately statistically independent of $\mathbf{bs}^{f1}(d+1) - \mathbf{bs}^{f2}(d+1)$, with 1-day lagged rank correlations ranging from 0.08 for 1-day forecasts to 0.21 for 5-day forecasts. Thus, the data was judged to be amenable to a paired resampling strategy, with these distinct vector blocks of data for each day. The following process was then repeated 10,000 times. For each of the 123 days, a random uniform number between 0 and 1 was generated. If the number was greater than 0.5, $\mathbf{bs}^{f1}(d)$ was randomly selected for inclusion in sample 1, $\mathbf{bs}^{f2}(d)$ was selected for inclusion in sample 2, and vice versa if the number was less than or equal to 0.5. The vector of average Brier scores for samples *s*1 and *s*2 were then calculated, $\overline{\mathbf{bs}}^{s1}$ and $\overline{\mathbf{bs}}^{s2}$. The *BSS* for samples 1 and 2 were generated via eq. (1), and the difference between the *BSS*s for the two samples was noted. The confidence intervals are the 5th and 95th percentiles of the difference between the *BSS*s of the two samples from the distribution generated through the 10,000 iterations.

These block bootstrap confidence intervals should be regarded as approximations. An assumption underlying this process is that there were 123 independent data samples. However, $\mathbf{bs}^{f1}(d)$ and $\mathbf{bs}^{f1}(d+1)$ were slightly correlated as discussed above, especially for the longer-lead forecasts, which will contribute a slight over-estimate of the effective sample size and thus underestimate of the confidence interval. On the other hand, data from grid points across the CONUS were aggregated in this procedure and thereafter considered as a single block. In reality there may be far more than one independent sample spanning the CONUS, thus leading to an under-estimate of sample size and consequent overestimate of the confidence interval in this approach. Note also that for simplicity of presentation, the skill diagrams will show only one set of confidence intervals, e.g., between NCEP and ECMWF forecasts. Slightly smaller confidence intervals could be expected were they computed using ECMWF and CMC forecasts, given their more similar skills.

In order to make sure that the *CRPSS* did not excessively reflect the skill of the climatologically wet grid points, an alternative to the standard method of *CRPSS* calculation was followed here. This method is analogous to how *CRPSS* would be computed if the forecasts were of probabilities of exceedance of various quantiles. An example of such a forecast product expressed in quantiles are NCEP/CPC's 6-10 day and 8-14 day forecasts (e.g., http://www.cpc.ncep.noaa.gov/products/predictions/610day/), which provide probabilities of below-normal/near-normal/above-normal temperature and precipitation, i.e., probabilities for the < 1/3 and ≥ 2/3 quantiles. In this alternative method of calculation, the *CRPS* at a given grid point was *not* computed by integrating differences between observed and forecast cumulative distribution functions (CDFs) *over a range of precipitation values* (the standard method). Instead, the differences between observed and forecast CDFs were integrated *over the percentiles of the CDF*, which were determined separately for each model grid point and each month. Specifically, given n_d case days, for the $s = 1, \ldots, n_\mathrm{d} \times n_\mathrm{s}$ samples, let $\mathbf{q}^s = \left[q_1^s, \ldots, q_{20}^s\right]$ be the 20-dimensional vector of the precipitation quantiles associated with the 2.5th, 7.5th, …, 97.5th percentiles of the climatological *CDF* for that point and that month. The average forecast $CRPS_f$ was determined by integrating in steps of 5 percent:

$$CRPS_f = \frac{\sum_{s=1}^{n_d \times n_s} \cos(\phi_s) \sum_{iq=1}^{20} 0.05 \times \left[F^s\left(q_{iq}^s\right) - O^s\left(q_{iq}^s\right) \right]^2}{\sum_{s=1}^{n_d \times n_s} \cos(\phi_s)} \qquad (4)$$

,

where $F^s\left(q_{iq}^s\right)$ represents the forecast's *CDF* for the *s*th sample evaluated at the q_{iq}^s quantile, and $O^s\left(q_{iq}^s\right)$ represents the same, but for the observed (analyzed). ϕ_s is the latitude of the grid box, the cosine factor accounting for latitudinal variations in grid box size. For raw ensemble guidance, $F^s\left(q_{iq}^s\right)$ was directly computed from the ensemble relative frequency. For example, if 5 of 20 members had values exceeding q_{iq}^s, then $F^s\left(q_{iq}^s\right)$= 0.75. For post-processed forecasts, $F^s\left(q_{iq}^s\right)$ was determined by numerical integration of eqs. (1) and (2). For the observed CDF, the analyzed state was assumed perfect, i.e., no analysis errors were incorporated, so the analyzed CDF was a Heaviside function, 0 at the quantiles less than the analyzed value, 1 at quantiles greater than or equal to the analyzed value. The *CRPS* of the climatological forecast, $CRPS_c$, was calculated as in eq. (4), but substituting the climatological CDF for the forecast CDF. Finally, the overall skill score was calculated as *CRPSS* = 1. - $CRPS_f$ / $CRPS_c$. As with the *BSS*, a paired block bootstrap approach was used to estimate the confidence intervals.

Two other common verification statistics were also used, root-mean-square (RMS) errors, and bias, the average forecast divided by the average analyzed amount.

3. **Results.**

a. Properties of forecasts from the individual centers.

Before considering the multi-model and ECMWF reforecast-calibrated forecast properties, let us consider the verification of PQPFs from the individual centers. Figure 2 shows > 1 mm (24 h)$^{-1}$ and > 10-mm (24 h)$^{-1}$ event *BSS* and *CRPSS*. ECMWF generally produced the most skillful raw precipitation PQPFs. Depending on the metric, either NCEP or UKMO produced the least skillful forecasts.

Interestingly, though UKMO forecasts appeared to be generally more skillful than NCEP forecasts in *BSS*, they appeared to be consistently worse in *CRPSS*. This was a consequence of the *CRPSS* verification algorithm as implemented here, which attempted to equally weight the *CRPSS* at all grid points, irrespective of whether the climatological event probability was extremely high or extremely low. The conventionally calculated *CRPSS* is dominated by the performance of the forecasts in the climatologically wet areas (Hamill and Juras 2006). There is inherently greater climatological variance of precipitation for the wet regions, and associated with this there are generally much larger *CRPS* values than in dry regions. Consequently, when evaluated over many grid points, the conventionally calculated *CRPS* and hence the *CRPSS* are dominated by the performance at the wetter points. Figure 3 shows maps of the day +3 *CRPSS* scores (see the online appendix for *CRPSS* maps for the other lead times). The UKMO forecasts had negative skill in the extremely dry regions of the western US. The RMS errors of the ensemble-mean forecasts in the dry regions of all the models were very small and relatively similar (Fig. 4a; for other lead times, see the online appendix). However, the UKMO forecasts exhibited a large moist bias in the climatologically dry regions (Fig. 4b), which resulted in a very large over-forecast of probabilities and poor skill for those points.   This was apparently due to a drizzle over-forecast bias in that version of the UKMO's forecast model (D. Barker, personal communication, 2011).

Figure 4b also illustrates some other interesting characteristics of the ensemble systems.   NCEP over-forecasted rainfall for the grid points and dates where the climatological probability was already quite high. CMC forecasts were also biased, exhibiting a moist bias at the lowest climatological probabilities but dry biases for most of the rest of the larger climatological probabilities. ECMWF forecasts were the least biased, with a moderate over-forecast bias at the low climatological probabilities.

Figure 5 provides reliability diagrams of day +3 forecasts of the > 10-mm (24 h)$^{-1}$ event. Other reliability diagrams for other lead times and for the > 1-mm (24 h)$^{-1}$ event are available in the online appendix. CMC forecasts were generally the most reliable, though they were not as sharp as the ECMWF forecasts and hence had a lower *BSS*. UKMO and NCEP forecasts were much less reliable, though NCEP forecasts were slightly sharper than the others. ECMWF 50-member forecasts were slightly more reliable and skillful than their 20-member subset.

In subjective analyses of individual forecasts, it appeared that several of the forecast models had subtle systematic northward biases in the northern central US. Figure 6 shows the 10-mm observed contour and the 0.5 probability contour for the > 10-mm (24 h)$^{-1}$ event from the day +3 ECMWF forecasts. Here, the 25 cases with the largest areal coverage of observed precipitation between 105° and 80° west longitude and 35° and 50° north latitude were chosen. Similar plots for the other forecast models are included in the online appendix.

b. Properties of multi-model and statistically post-processed forecasts.

Before considering verification scores, consider first two actual forecast cases, presented in Figs. 6 and 7, showing probabilities from the 20-member ensembles and from the 80-member multi-model ensemble. The first case, covering the 24-h period ending 00 UTC 21 July 2010, illustrates that sometimes the forecast models could be overly similar to each other. Here all the forecast precipitation areas were significantly north of the observed area. A multi-model forecast would not be expected to provide much benefit in such a situation. Figure 7 shows the same type of plot, but for 24-h period ending 00 UTC 7 August 2010. Here the multi-model forecast provided some improvement. On this day the CMC and UKMO areas of high probabilities were too far north, the NCEP area too far south, but the higher probabilities in the multi-model forecasts were more coincident with the analyzed regions exceeding 10 mm. Most of the area with greater than 10 mm in the analysis were covered by nonzero multi-model probabilities. More generally, when there was some diversity of positions in the multi-model forecasts, this often allowed the forecast to avoid being inappropriately sharp.

Figure 9 provides *BSS*s and *CRPSS* for the multi-model and the post-processed forecasts. For the light precipitation forecasts (> 1.0 mm (24 h)$^{-1}$, Fig. 9a), the multi-model forecasts improved the skill by approximately +1 day relative to ECMWF at the earliest lead times; a +2 day multi-model forecast could now be made as skillfully as a +1 day ECMWF forecast. The improvement in skill was a more modest $\sim$ +0.3 days at the longer forecast lead times. The calibrated multi-model forecast product improved skill over the basic multi-model forecast by a tiny amount at day +1 but degraded the skill after day +3. This is consistent with previous results; at the longer lead times, the growth of errors makes it more difficult to differentiate the model bias from the chaotically induced errors with short training data sets (Hamill et al. 2004). The improvement from reforecast-based post-processing over the raw ECMWF system was much smaller than the improvement from single to multi-model and was even slightly negative at the day +5 lead. Reasons for the less impressive performance of reforecast calibration than in previous studies will be discussed at the end of this section.

More impressive increases in skill were evident for the > 10-mm $(24\ \mathrm{h})^{-1}$ event. Both the reforecast-based calibration and the multi-model approach increased forecast skill by an equivalent of up to +2 days of additional lead time. Again, the calibration of the multi-model forecasts provided modest improvement at the early leads and degradation at the longer leads relative to the unprocessed multi-model.

Measured in *CRPSS*, the multi-model forecasts produced the most skillful forecasts, exceeding the skill of reforecast-calibrated ECMWF forecasts by a small amount. Consider now *where* the forecasts were improved or degraded by the various approaches. Figure 10 provides maps of the day +3 *CRPSS*; maps for other lead times are in the online appendix. The patterns of multi-model skill are rather similar to those of the most skillful ensemble system, ECMWF (Fig. 3a). The reforecast-calibrated ECMWF forecasts appear to have increased the skill most notably in the driest regions of the western US.

Figure 11 shows day +3 > 10-mm (24 h)$^{-1}$ event reliability diagrams for the multi-model, the calibrated multi-model, and reforecast-calibrated ECMWF PQPFs. The raw multi-model PQPFs were slightly more reliable than any of the PQPFs from the individual centers (Fig. 5) and retained a slight over-forecast bias at the higher probabilities. The improvements in reliability were more substantial than for the > 1-mm (24 h)$^{-1}$ event; see diagrams in the online appendix. The reforecast-calibrated PQPFs exhibited a slight under-forecast bias and were not as sharp as those from the multi-model forecasts. Was this due to some inhomogeneity between the 2002-2009 training data and the 2010 real-time forecasts? Figure 12 shows that there were fewer large forecast busts in 2010 than there were in 2002 or 2006. When the regression analysis from 2002-2009 data was applied to correct the 2010 forecasts, the assumption was that the 2010 forecasts would be equally unskillful. In fact they were better, and as a consequence the post-processed forecasts were less sharp than they could have been. Though it was not attempted here, it might be possible to apply ad-hoc corrections to the training data to improve the regression analysis. Perhaps a slight adjustment of the training data ensemble mean toward the analyzed data would make its accuracy more closely resemble that of the 2010 data, sharpening and making the ELR forecasts more reliable and skillful.

Figure 13 shows the multi-model areal coverage of the 0.5 probability contours for the > 10 mm (24 h)$^{-1}$ event for selected cases; these should be compared with Fig. 6 for ECMWF-only PQPFs. Figure 14 also shows the areal coverage, but for reforecast-calibrated ECMWF PQPFs.   The areal coverage was only slightly smaller for the multi-model PQPFs than it was for the ECMWF PQPFs, illustrating that the multi-model forecasts did not lose a tremendous amount of sharpness (coverage of greater than 0.5 probability being a proxy for sharpness here). In comparison, the reforecast-calibrated PQPFs in Fig. 14 show a marked decrease in the areal coverage; many grid points with probability $p > 0.5$ in the raw ECMWF PQPF had $p < 0.5$ after calibration. Figures 15 and 16 show for the cases plotted in Figs. 7 and 8 a bit more detail on what happened with typical multi-model and reforecast-calibrated PQPFs. The multi-model forecasts retained their sharpness, but not always desirably so. For example, in Fig. 15, the multi-model forecasts retain relatively high probabilities in eastern Iowa and northern Illinois, whereas the analyzed area was displaced further south. The reforecast-calibrated PQPFs decreased the areal coverage of high probabilities, appropriately so in this case, reducing the false alarms. However, as seen in inspection of Figs. 13-14, there were many cases when the sharpness retained in the multi-model forecasts was desirable.

The results exhibited here with reforecast calibration were not as impressive as they have been in previous studies, e.g., Hamill and Whitaker (2006) and Hamill et al. (2008). There are at least four reasons for this. First, the training data was not as accurate as the real-time data in this application (Fig. 12), and this inhomogeneity degraded the regression analysis. This may have been due to less accurate initial conditions (ERA-Interim for the reforecast, operational 4D-Var for the real-time forecasts) and because the reforecast ensemble was initialized with perturbations that were constructed with approximations different from those in the real-time forecasts (section 2b). The second reason is that gratifying improvements have been made to models and EPSs so that they produce more skillful and reliable forecasts than they did even in the recent past; it's tougher to improve upon ECMWF's 2010's model output than its 2005 model output. The third reason is that even with the use of ECMWF's reforecasts, there really was a limited training data set in this study, here due to the unavailability of precipitation analyses prior to 2002 and the unavailability of reforecast data more frequently than once per week. The fourth reason is that in prior studies, the ensemble forecasts (at coarse resolution) were evaluated against analysis data at finer resolution, so that the reforecast calibration process was also producing a statistical downscaling. This point is worth keeping in mind when considering the relative merits of reforecast calibration vs. multi-model approaches. If the desired output is forecast data at the grid scale, multi-models may have substantial appeal. If the desired output is point data or high-resolution gridded data, the statistical downscaling is more straightforward when reforecasts are used.

Overall, the impressive skill improvements provide evidence for the merit of both multi-model ensemble and reforecast approaches. Should other forecast centers share precipitation ensemble data, large gains in probabilistic precipitation forecast skill are possible for little more than the cost of data transmission and storage. Alternatively, should any one center produce and utilize reforecasts, they can improve their own forecasts significantly, assuming a comparably long time series of observations or analyses are available. The improvement here noted with reforecasts may have also been modest because the training data was limited on account of a short time series of analyses, dating back to only 2002; only around 40% of the available reforecast data was used.

4. Conclusions.

This article examined probabilistic multi-model weather forecasts of precipitation over the CONUS and the relative advantages and disadvantages of these forecasts when compared to statistically post-processed ECMWF forecasts. 20-member forecasts were extracted from the ECMWF, NCEP, UKMO, and CMC global ensemble systems at 1-degree resolution between June and October 2010. Daily 24-h accumulated probabilistic precipitation forecasts were generated from the subsequent 80-member ensemble for lead times of +1 to +5 days and compared to gridded precipitation analyses. Two statistically post-processed products were also evaluated, the first being multi-model forecasts that were adjusted using extended logistic regression and that were trained on the previous 30 days of forecasts and analyses. The second was ECMWF forecasts, which were statistically adjusted using forecast/analysis data for the period 2002-2009, the time period when both reforecasts and analyses were available.

Considering first the skill of forecasts from the individual EPSs, ECMWF forecasts generally were the most skillful in terms of Brier skill scores and the continuous ranked probability skill score. CMC forecasts were the most reliable but the least sharp, while NCEP and UKMO forecasts were more sharp but less reliable.

Multi-model probabilistic forecast products were substantially more skillful than the best of the individual centers' probabilistic forecasts. The improvement was approximately an extra +0.5 to +1 day of forecast lead time for light precipitation events and as much as +2 days for heavier precipitation events. The reforecast-calibrated ECMWF forecasts exhibited more skill and reliability improvement at the > 10-mm (24 h)$^{-1}$ event as they did at the > 1-mm (24 h)$^{-1}$ event. Relative to the multi-model forecasts, the reforecast-calibrated skills were similar for the > 10 mm (24 h)$^{-1}$ event, but the reforecast-calibrated was more reliable while the multi-model was sharper.

The results exhibited here with reforecast calibration were not as impressive as they have been in previous studies. There were at least four reasons for the lessened improvement of reforecast calibration here. First, the reforecast training data was shown to be not as accurate as the real-time data in this application. Second, gratifying improvements have been made to models and EPSs in the last few years; it's tougher to improve upon ECMWF's 2010's model output than its 2005 model output. Third, limited training data set was available for this study. Fourth, prior studies were performed at higher resolution and produced a statistical downscaling that the coarser raw forecasts could not accomplish.

I was pleasantly surprised by the magnitude of skill improvements demonstrated here from multi-model ensembles, improvements which were larger than those seen with 2-meter temperatures (Hagedorn et al. 2011). From our own experience, however, I recommend some caution against broadly generalizing these results to any multi-model ensemble system. This study examined a combination of data from four mature EPSs based on mature models and assimilation systems. Each center's system has been refined through the collective efforts of hundreds if not thousands of person-years of research and development. A combination of less developed EPSs may not provide nearly the same gratifying result.

Nonetheless, these results demonstrate the potential value of multi-model ensembles. The THORPEX program, organized by the World Meteorological Organization, has promoted the concept of a multi-model based "Global Interactive Forecast System" (Bougeault et al. 2010), whereby the operational centers share data that will facilitate the production of multi-model products for high-impact weather events. This study provides additional evidence for the validity and the potential benefits of such a system. Currently several centers have restrictive data policies; full access to their data is reserved for paying customers, and those customers cannot thereafter share the data they purchased. Perhaps the approach embraced in the US and Canada will be followed by other centers worldwide, for the mutual benefit of all. In the US and Canada, the data is effectively free since the research, development, and production were funded by public taxpayer funds.

Finally, can we all have "the best of both worlds?" That is, will NWP centers both agree to share their ensemble data freely and internationally in real time, and will they produce reforecast data sets so that each model can be calibrated to remove systematic errors prior to their combination? There is evidence that such approaches will provide substantial benefit. The climate community is working on sharing multi-model information and hindcasts to facilitate the error correction for intra-seasonal and seasonal forecasts. For weather and weather-to-climate applications, there have also been successful demonstrations of multi-model calibrated forecasts (Vislocky and Fritsch 1995, Whitaker et al. 2006). NOAA is currently developing a new reforecast data set for its global ensemble prediction system, and I hope that other centers will be inspired to do so as well.

Acknowledgments

TIGGE data was obtained from ECMWF's TIGGE data portal. I thank ECMWF for the development of this portal software and for the archival of this immense data set. Florian Pappenberger of ECMWF was very helpful in extracting and pre-processing the reforecasts. Yan Luo of NCEP/EMC was helpful in obtaining the CCPA analysis data. Tom Galarneau of NCAR/MMM is thanked for providing an informal review. Publication of this study was funded by the NOAA THORPEX program. Three anonymous reviewers are thanked for their helpful recommendations to improve this article.

Appendix 1.

Here are additional details on the forecast models and ensemble systems used in this experiment.

a. NCEP

NCEP used the Global Forecast System (GFS) model in their ensemble system at T190L28 resolution. A lengthier description of the physical packages used in this model were described in Hamill et al. (2011). A description of the GFS model is available from the NCEP Environmental Modeling Center (EMC), with changes as of 2003 described at www.emc.ncep.noaa.gov/gmb/moorthi/gam.html.

The control initial condition around which the perturbed initial conditions were centered was produced by the T382 Global Statistical Interpolation (GSI) analysis (Kleist et al. 2009) at T384L64 resolution. Perturbed initial conditions were generated with the ensemble transform with rescaling technique of Wei et al. (2008). Stochastic perturbations were included, following Hou et al. (2008). More details on changes to the NCEP ensemble system can be found at http://www.emc.ncep.noaa.gov/gmb/yzhu/html/ENS_IMP.html.

b. Canadian Meteorological Centre

The CMC EPS used the Global Environmental Multiscale Model, a primitive equation model with a terrain-following pressure vertical coordinate. Further documentation on the GEM model can be found at http://collaboration.cmc.ec.gc.ca/science/rpn/gef_html_public/DOCUMENTATION/GENERAL/general.html and in Charron et al. (2010). The CMC ensemble system used a horizontal computational grid of 400x200 grid points, or approximately 0.9 degrees, and 28 vertical levels. The EnKF initial conditions were used, following Charron et al. (2010) and Houtekamer et al. (2009) and references therein. The 20 forecast ensemble members used a variety of perturbed physics; changing gravity wave drag parameters, land-surface process type, condensation scheme type, convection scheme type, shallow convection scheme type, mixing-length formulation, and turbulent vertical diffusion parameter. More details on these are provided at http://www.weatheroffice.gc.ca/ensemble/verifs/model_e.html.

c. European Centre for Medium-Range Weather Forecasts.

The ECMWF EPS used the ECMWF Integrated Forecast System (IFS) model, versions 36r2. Model resolution was T639L62 for both versions; details on the IFS are provided at www.ecmwf.int/research/ifsdocs/. The changes to the ensemble stochastic treatments in the 8 Sep 2009 implementation are described in Palmer et al. (2009). The ensemble was initialized with a combination of initial-time and evolved total-energy singular vectors (Buizza and Palmer 1995, Molteni et al. 1996, Barkmeijer et al. 1998, Barkmeijer et al. 1999, Leutbecher 2005) and utilized stochastic perturbations to physical tendencies. An overview of the ensemble system was provided in Buizza et al. (2007) and references therein. For consistency with the analysis of other EPSs, only the first 20 perturbed members were used here.

e. United Kingdom Met Office.

The UK Met Office (UKMO) ensemble system was "MOGREPS," the Met Office Global and Regional Ensemble Prediction System. TC track forecasts from this system came from its global component, which was described in Bowler et al. (2008, 2009). The global system was run at a resolution of 0.83° longitude and 0.55° latitude on a regular latitude-longitude grid. 70 vertical levels were employed (Tennant et al. 2011). Initial condition perturbations were generated from an implementation of the ensemble transform Kalman filter (Hunt et al. 2006, Bowler et al. 2009). The mean initial state was generated from the UKMO 4D-Var system (Rawlins et al. 2007). The model included a parameterization of one type of model uncertainty via its stochastic kinetic-energy backscatter scheme, following Shutts

(2005) and Tennant et al. (2011).

Appendix 2:

This appendix discusses the method used to augment the training sample size used in the regression analyses. Were only the data at the grid point of interest used for training, when calibrating using the multi-model ensemble using the past 30 days of forecasts and analyses, this would provide, of course, only 30 training samples. Older forecasts could be used, but precipitation biases are often seasonally dependent, so the older data may degrade the results despite augmenting the sample size. Also, with such a multi-model ensemble, the farther back into the past one seeks training data, the more likely it is that at least one of the models will have had a major upgrade and concomitant change in systematic error characteristics.

Despite ECMWF providing a multi-decadal reforecast, in practice the sample sizes were too small here, too. When using the 2002-2009 weekly, 5-member ECMWF reforecasts (including reforecast dates +/- 6 weeks around the week of interest), this provided a total of 13 weeks $\times$ 8 years = 104 samples. In both cases, these were relatively small samples to estimate four regression parameters, and especially for rare events such as heavy precipitation, experience has shown that larger training samples improved the regression analysis.

Hence, following the general philosophy demonstrated and discussed in Hamill et al. (2008) and inspired by the regionalization used in some Model Output Statistics algorithms (Lowry and Glahn 1976), the training data set for a particular grid point was augmented by finding 25 other grid points that had relatively similar climatological analyzed CDFs. Consider a particular location (λ, ϕ) at which we seek to augment the sample size, and another location $\left(\lambda^{s},\phi^{s}\right)$ we are considering as a location with suitable supplemental training data. Differences between the analyzed cumulative probabilities at $\left(\lambda,\phi\right)$ and $\left(\lambda^{s},\phi^{s}\right)$ were measured at the 1, 2.5, 5, 10, 25, and 50 mm (24 h)$^{-1}$ amounts and then weighted by similar respective factors of [1, 2.5, 5, 10, 25, 50]. That is, a cumulative probability difference of 0.1 at 1 mm and 0.1/50 at 50 mm were judged to have the same weighted difference (this approach is admittedly somewhat arbitrary, and testing found that the overall calibration results were relatively insensitive to the details of this assumption). The maximum weighted difference at any of the possible precipitation amounts was then noted for this $\left(\lambda^{s},\phi^{s}\right)$. Having evaluated the maximum of the weighted differences all the grid points less that 8 grid points distant from the grid point of interest $\left(\lambda,\phi\right)$, the 25 grid points with the smallest weighted differences were identified, and the training sample for $\left(\lambda,\phi\right)$ was augmented by the forecasts-analysis pairs at these locations. This approach increased sample size, but it’s possible that the forecast bias might have been different at the supplemental locations, and hence not an unalloyed benefit. For more discussion of this, see Hamill et al. (2008, section 3a).

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids. *Weather and Forecasting*, **18,** 918-932.

Bao, L., T. Gneiting, E. P. Grimit, P. Guttorp, and A. E. Raftery, 2010: Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction. *Monthly Weather Review*, **138,** 1811-1821.

Barkmeijer, J., F. Bouttier, and M. Van Gijzen, 1998: Singular vectors and estimates of the analysis-error covariance metric. *Quarterly Journal of the Royal Meteorological Society*, **124,** 1695-1713.

Barkmeijer, J., R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **125,** 2333-2351.

Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multi-physics representations. *Monthly Weather Review*, **139,** 1972-1995.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A Spectral Stochastic Kinetic Energy Backscatter Scheme and Its Impact on Flow-Dependent Predictability in the ECMWF Ensemble Prediction System. *Journal of the Atmospheric Sciences*, **66,** 603-626.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, **91,** 1059-1072.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **135,** 767-776.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **134,** 703-722.

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Quarterly Journal of the Royal Meteorological Society*, **133,** 681-695.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **125,** 2887-2908.

Buizza, R., and T. N. Palmer, 1995: The Singular-Vector Structure of the Atmospheric Global Circulation. *Journal of the Atmospheric Sciences*, **52,** 1434-1456.

Candille, G., 2009: The Multiensemble Approach: The NAEFS Example. *Monthly Weather Review*, **137,** 1655-1665.

Charron, M., G. r. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System. *Monthly Weather Review*, **138,** 1877-1901.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137,** 553-597.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A*, **57,** 234-252.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS Uncertainty Estimates in an Ensemble Framework. *Monthly Weather Review*, **137,** 246-268.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133,** 1098-1118.

Hagedorn, R., 2008: Using the ECMWF reforecast data set to calibrate EPS reforecasts. *ECMWF Newsletter*, **117,** 8-13.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2011: Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **accepted pending revision; available from martin.leutbecher@ecmwf.int.**

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A*, **57,** 219-233.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Monthly Weather Review*, **136,** 2608-2619.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Monthly Weather Review*, **136,** 2620-2632.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, **132,** 2905-2923.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, **134,** 3209-3229.

——, 2007: Ensemble Calibration of 500-hPa Geopotential Height and 850-hPa and 2-m Temperatures Using Reforecasts. *Monthly Weather Review*, **135,** 3273-3280.

Hamill, T. M., J. S. Whitaker, M. Fiorino, and S. G. Benjamin, 2011: Global Ensemble Predictions of 2009's Tropical Cyclones Initialized with an Ensemble Kalman Filter. *Monthly Weather Review*, **139,** 668-688.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An Important Dataset for Improving Weather Predictions. *Bulletin of the American Meteorological Society*, **87,** 33-46.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review*, **132,** 1434-1447.

Higgins, R. W., J. E. Janowiak, and Y.-P. Yao, 1996: A Gridded Hourly Precipitation Data Base for the United States (1963-1993). *NCEP/Climate Prediction Center ATLAS No. 1, U. S. DEPARTMENT OF COMMERCE, National Oceanic and Atmospheric Administration, National Weather Service*.

Hou, D., Z. Toth, Y. Zhu, and Y. Yang, 2008: Impact of a Stochastic Perturbation Scheme on NCEP Global Ensemble Forecast System. *Proceedings, 19th AMS conference on Probability and Statistics. New Orleans, LA, 20-24 Jan. 2008.*

Houtekamer, P. L., and H. L. Mitchell, 2005: Ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, **131,** 3269-3289.

Houtekamer, P. L., H. L. Mitchell, and X. Deng, 2009: Model Error Representation in an Operational Ensemble Kalman Filter. *Monthly Weather Review*, **137,** 2126-2143.

Hunt, B., E. Kostelich, and I. Szunyogh, 2006: Efficient Data Assimilation for Spatiotemporal Chaos: a Local Ensemble Transform Kalman Filter.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, L. Raynaud, 2010: Ensemble of data assimilations at ECMWF, 48 pp.

Iversen, T., A. Deckmyn, C. Santos, K. A. I. Sattler, J. B. Bremnes, H. Feddersen, and I.-L. Frogner, 2011: Evaluation of 'GLAMEPS'—a proposed multimodel EPS for short range forecasting. *Tellus A*, **63,** 513-530.

Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, **135,** 777-794.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP Global Data Assimilation System. *Weather and Forecasting*, **24,** 1691-1705.

Krishnamurti, T. N., and Coauthors, 2000: Multimodel Ensemble Forecasts for Weather and Seasonal Climate. *Journal of Climate*, **13,** 4196-4216.

Leutbecher, M., 2005: On Ensemble Prediction Using Singular Vectors Started from Forecasts. *Monthly Weather Review*, **133,** 3038-3046.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. *Preprints, 19th Conf. on Hydrology, American Meteorological Society, San Diego, CA 9-13 January 2005, Paper 1.2.* .

Lorenz, J., H. Rauhut, F. Schweitzer, and D. Helbing, 2011: How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, **108,** 920-925.

Lowry, D. A., and H. R. Glahn, 1976: An Operational Model for Forecasting Probability of Precipitation - PEATMOS PoP. *Monthly Weather Review*, **104,** 221-232.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122,** 73-119.

Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quarterly Journal of the Royal Meteorological Society*, **128,** 361-384.

Palmer, T. N., and Coauthors, 2009: Stochastic parameterization and model uncertainty. *ECMWF Tech Memo 589*.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133,** 1155-1174.

Rawlins, F., and Coauthors, 2007: The Met Office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, **133,** 347-362.

Roulin, E., and S. Vannitsem, 2011: Post-processing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review*, **139,** Available from Emmanuel.Roulin@meteo.be.

Schmeits, M. J., and K. J. Kok, 2010: A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts. *Monthly Weather Review*, **138,** 4199-4211.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, **131,** 3079-3102.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, **135,** 3209-3220.

Stensrud, D. J., and N. Yussouf, 2003: Short-Range Ensemble Predictions of 2-m Temperature and Dewpoint Temperature over New England. *Monthly Weather Review*, **131,** 2510-2524.

Tennant, W. J., G. J. Shutts, A. Arribas, and S. A. Thompson, 2011: Using a Stochastic Kinetic Energy Backscatter Scheme to Improve MOGREPS Probabilistic Forecast Skill. *Monthly Weather Review*, **139,** 1190-1206.

Vannitsem, S., and C. Nicolis, 2008: Dynamical Properties of Model Output Statistics Forecasts. *Monthly Weather Review*, **136,** 405-419.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved Model Output Statistics Forecasts through Model Consensus. *Bulletin of the American Meteorological Society*, **76,** 1157-1164.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a Short-Range Multimodel Ensemble System. *Monthly Weather Review*, **129,** 729-747.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, **60,** 62-79.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, **134,** 241-260.

Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving Week-2 Forecasts with Multimodel Reforecast Ensembles. *Monthly Weather Review*, **134,** 2279-2284.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences (2nd Ed.).* Academic Press, 627. pp.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, **16,** 361-368.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, **135,** 2379-2390.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging. *Monthly Weather Review*, **135,** 1364-1385.

Yussouf, N., and D. J. Stensrud, 2007: Bias-Corrected Short-Range Ensemble Forecasts of Near-Surface Variables during the 2005/06 Cool Season. *Weather and Forecasting*, **22,** 1274-1286.

Figure captions

Figure 1: Illustration of the process for determining precipitation classes used in the calculation of *BSS*. (a) Climatological probability of > 1-mm $24h^{-1}$ precipitation as determined from Stage-IV data for September 2002-2009. (b) Climatological class assigned to each grid point for September, 1-mm (24 h)$^{-1}$ event.

Figure 2: Brier skill scores of various forecasts for the (a) > 1-mm (24 h)$^{-1}$ event, (b) > 10-mm (24 h)$^{-1}$ event, and (c) continuous ranked probability skill scores, all as a function of forecast lead time. Error bars denote confidence intervals, the 5th and 95th percentiles of a paired block bootstrap between ECMWF and NCEP forecasts.

Figure 3: Maps of average *CRPSS* for day +3 forecasts for (a) ECMWF, (b) NCEP, (c) UKMO, and (d) CMC.

Figure 4: (a) RMS errors, and (b) bias for day +3 forecasts, each as a function of the climatological probability of greater than 1-mm (24 h)$^{-1}$. Light grey bars in panel (a) denote the relative frequency of each climatological probability.

Figure 5: Reliability diagrams for day +3 forecasts for the > 10-mm (24 h)$^{-1}$ event. (a) ECMWF, (b) NCEP, (c) CMC, and (d) UKMO. The dark line on each is the 20-member reliability curve. The lighter grey line on panel (a) is the reliability for the full 50-member ensemble. The inset histogram bars show the relative frequency of usage for each probability bin. The black lines on the inset are the relative frequency of usage for the climatological

distribution across all the sample points. The grey dots on the inset histogram of panel (a) are the relative frequency of usage for the ECMWF full 50-member ensemble.

Figure 6: Analyzed > 10-mm (24 h)$^{-1}$ precipitation boundary (black line) and area exceeding 10 mm (grey shading) for 25 cases with the largest areal coverage of greater than 10 mm in the upper Midwest US. Red lines indicate the 0.5 probability contour from the ECMWF ensemble for the day +3 forecasts of > 10 mm (24 h)$^{-1}$.

Figure 7: (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010. 10-mm (24 h)$^{-1}$ contour is denoted by the thick black line. (b) Probability of greater than 10 mm (24 h)$^{-1}$ for day +3 forecast from the ECMWF ensemble for the same period. The analyzed 10-mm contour from panel (a) is repeated. (c) as in (b), but for NCEP. (d) CMC, (e) UK Met Office, and (f) multi-model combination.

Figure 8: As in Fig. 7, but for 24-h period ending 00 UTC 8 August 2010.

Figure 9: Brier skill scores of various forecasts for (a) > 1-mm (24 h)$^{-1}$ event, and (b) > 10-mm (24 h)$^{-1}$ event, and (c) continuous ranked probability skill scores, all as a function of forecast lead time. "Multi-model/cal" refers to forecasts from the multi-model, calibrated using ELR. "ECMWF/reforecast" refers to ECMWF forecasts calibrated using ELR and the reforecast data set. Error bars denote confidence intervals, the 5th and 95th percentiles of a paired block bootstrap between ECMWF and NCEP forecasts.

Figure 10: Maps of *CRPSS* for day +3 forecasts for (a) multi-model, (b) multi-model with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.

Figure 11: As in Fig. 3, reliability diagrams for day +3 forecasts at the > 10-mm (24 h)$^{-1}$ event. (a) Multi-model forecasts, (b) multi-model with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.

Figure 12: A histogram of the absolute errors of day +3 ensemble-mean precipitation forecasts for the 2002 and 2006 reforecasts and for the 2010, 20-member real-time ensemble.

Figure 13: As in Fig. 6, but for multi-model forecasts.

Figure 14: As in Fig. 6, but for reforecast-calibrated ECMWF forecasts.

Figure 15: (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010. 10-mm contour is denoted by the thick black line. (b) Probability of greater than 10 mm (24 h)$^{-1}$ for day +3 forecast from the ECMWF ensemble for the same period. (c) as in (b), but for multi-model ensemble, and (d) as in (b), but for reforecast-calibrated ECMWF ensemble.

Figure 16: As in Fig. 15, but for the 24-h period ending 00 UTC 8 August 2010.

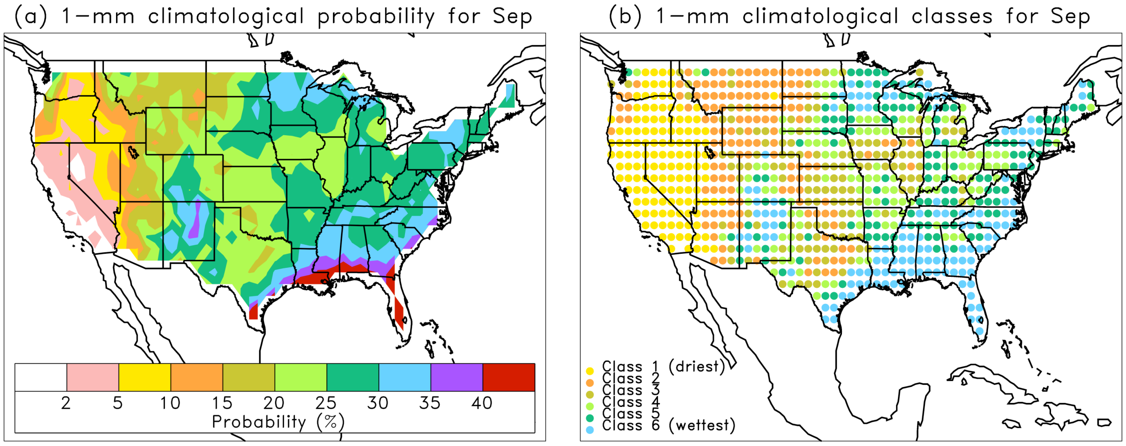


Figure 1: Illustration of the process for determining precipitation classes used in the calculation of *BSS*. (a) Climatological probability of > 1-mm 24h^{-1} precipitation as determined from Stage-IV data for September 2002-2009. (b) Climatological class assigned to each grid point for September, 1-mm (24 h)$^{-1}$ event.

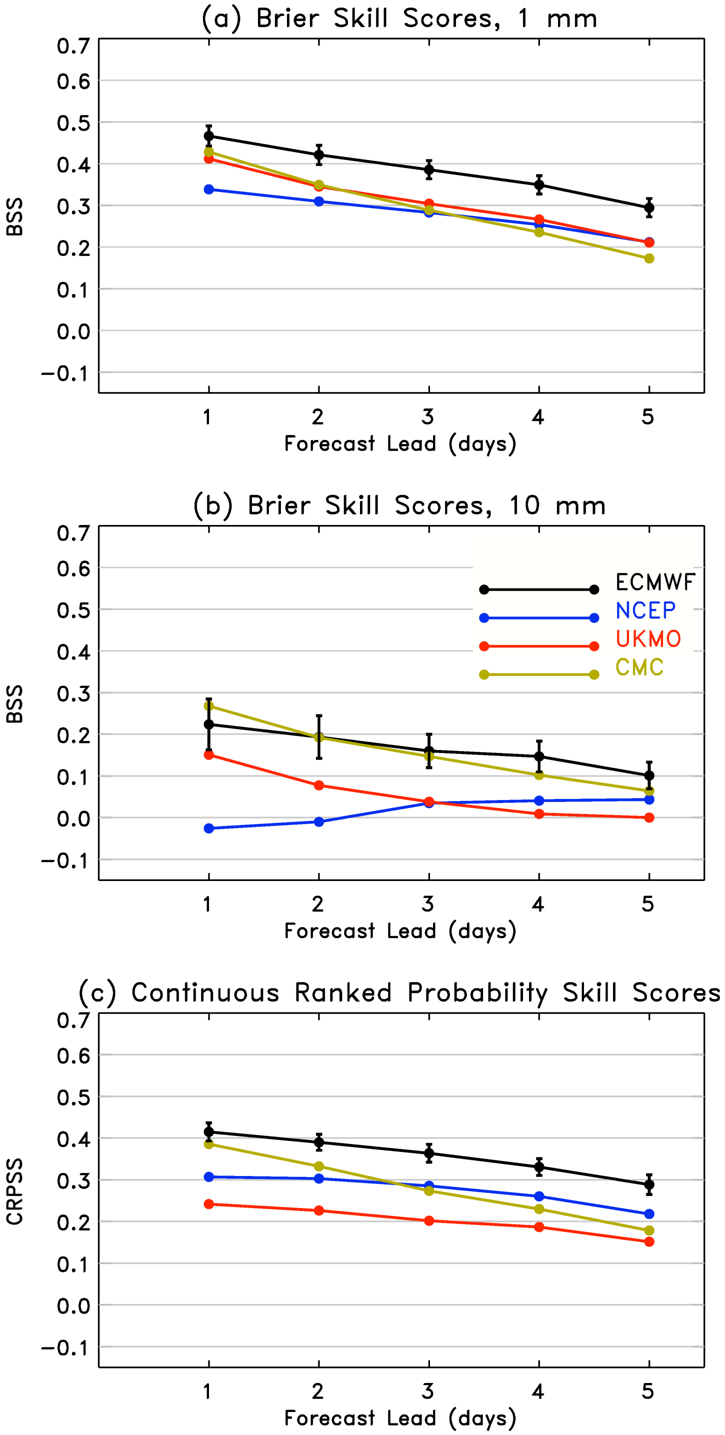


Figure 2: Brier skill scores of various forecasts for the (a) > 1-mm (24 h)$^{-1}$ event, (b) > 10-mm (24 h)$^{-1}$ event, and (c) continuous ranked probability skill scores, all as a function of forecast lead time. Error bars denote confidence intervals, the 5th and 95th percentiles of a paired block bootstrap between ECMWF and NCEP forecasts.

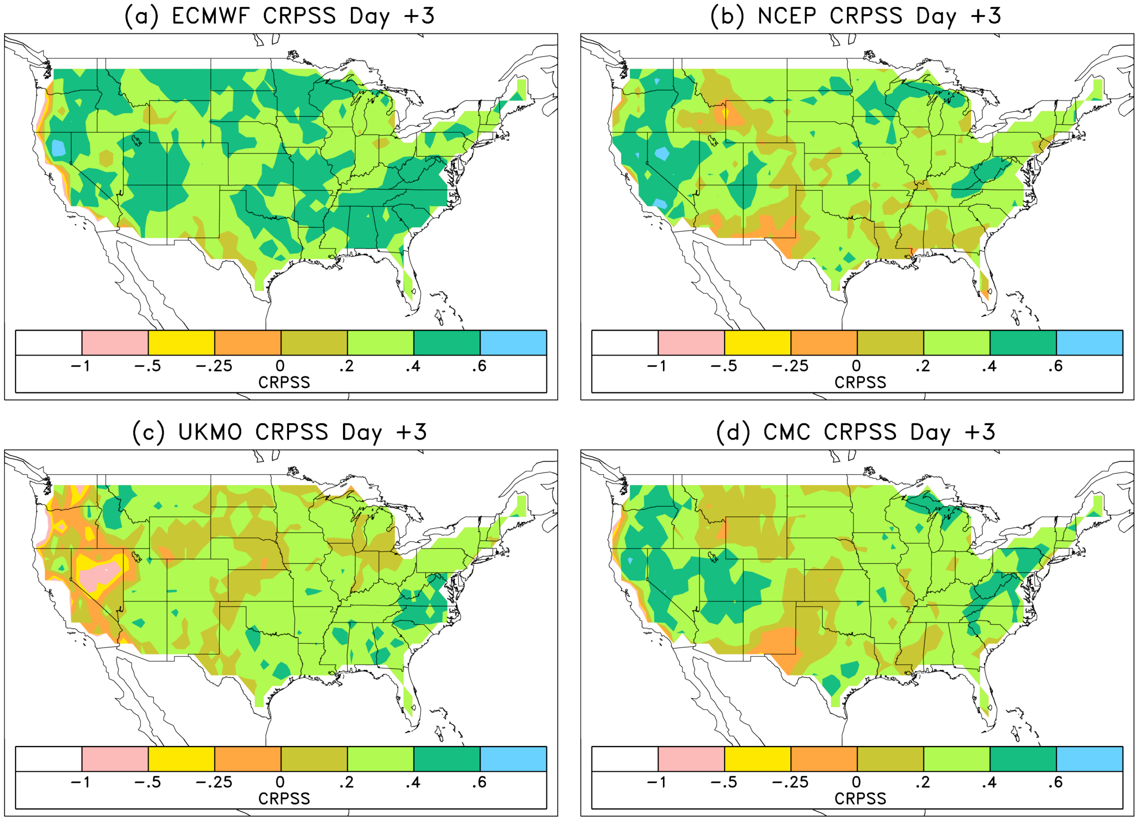


Figure 3: Maps of average *CRPSS* for day +3 forecasts for (a) ECMWF, (b) NCEP, (c) UKMO, and (d) CMC.

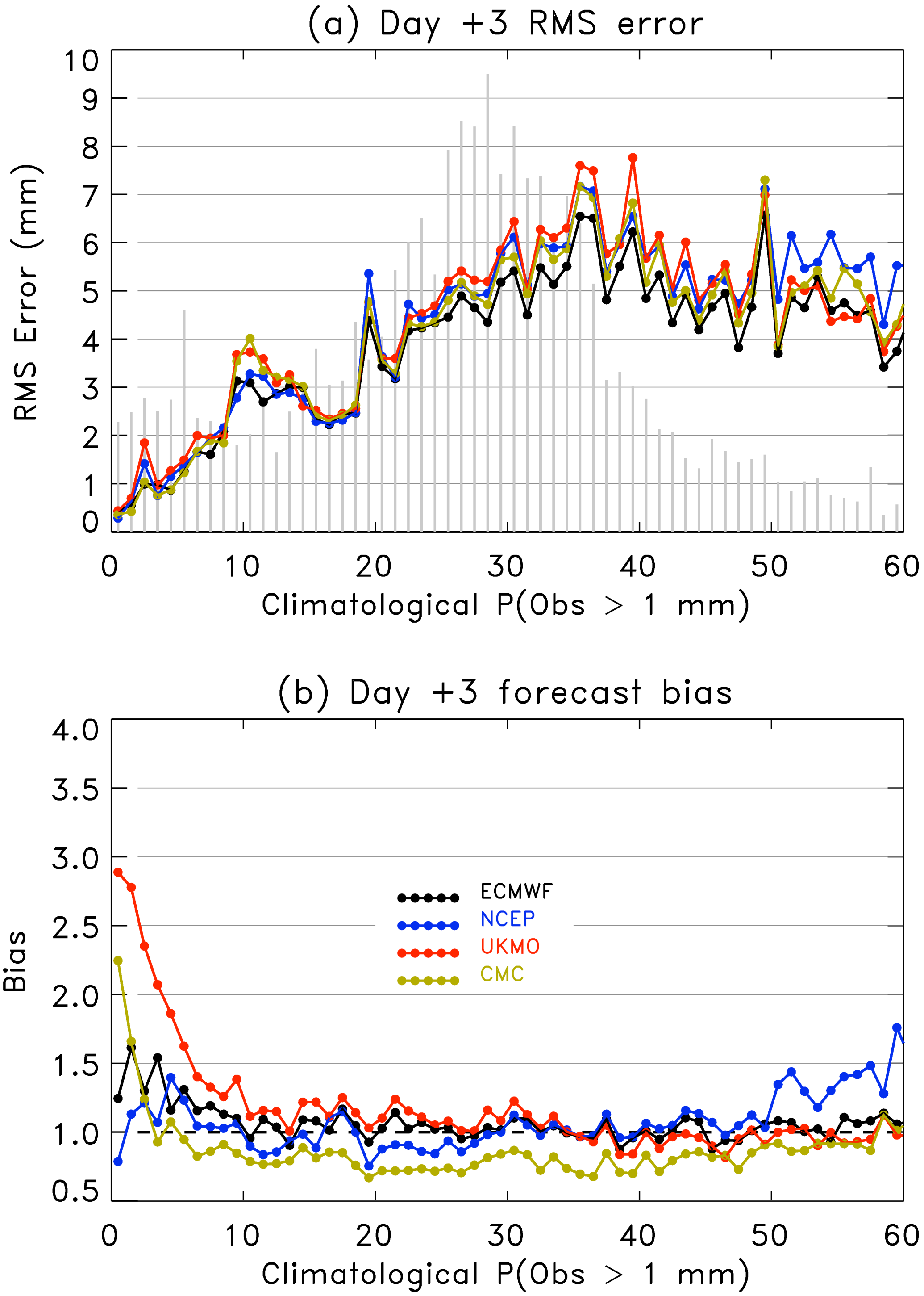


Figure 4: (a) RMS errors, and (b) bias for day +3 forecasts, each as a function of the climatological probability of greater than 1-mm (24 h)$^{-1}$. Light grey bars in panel (a) denote the relative frequency of each climatological probability.

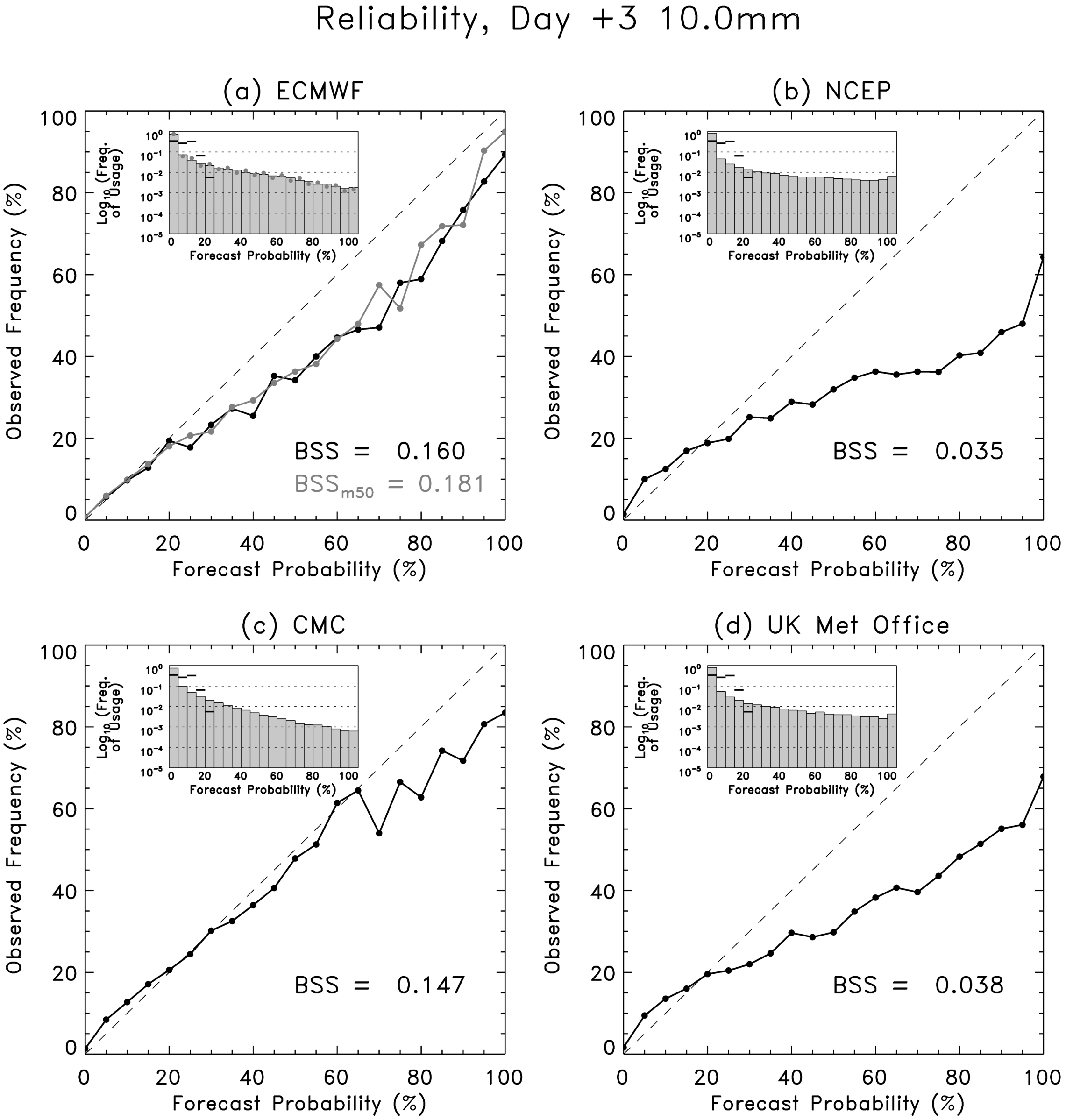


Figure 5: Reliability diagrams for day +3 forecasts for the > 10-mm (24 h)$^{-1}$ event. (a) ECMWF, (b) NCEP, (c) CMC, and (d) UKMO. The dark line on each is the 20-member reliability curve. The lighter grey line on panel (a) is the reliability for the full 50-member ensemble. The inset histogram bars show the relative frequency of usage for each probability bin. The black lines on the inset are the relative frequency of usage for the climatological distribution across all the sample points. The grey dots on the inset histogram of panel (a) are the relative frequency of usage for the ECMWF full 50-member ensemble.

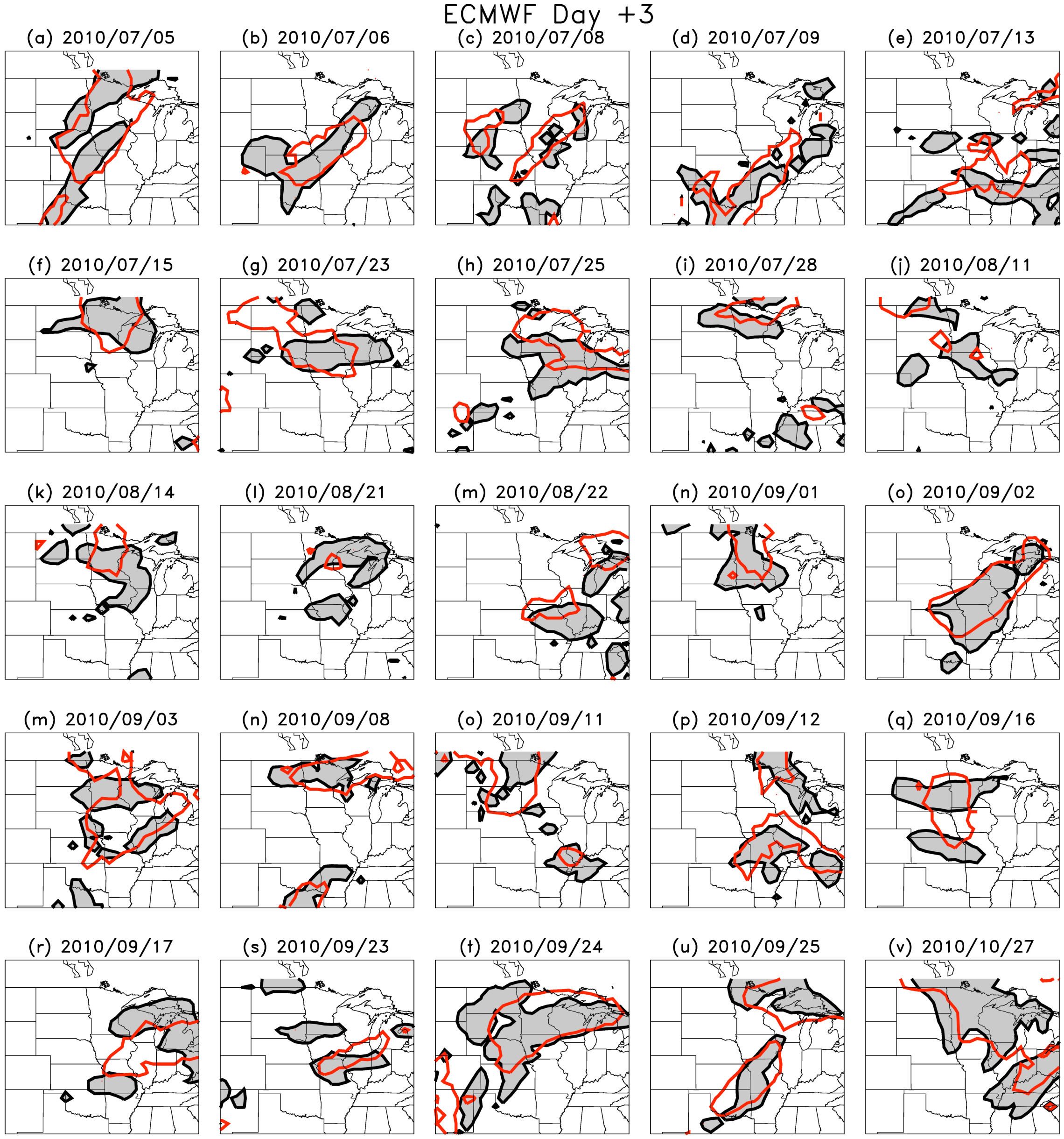


Figure 6: Analyzed > 10-mm (24 h)$^{-1}$ precipitation boundary (black line) and area exceeding 10 mm (grey shading) for 25 cases with the largest areal coverage of greater than 10 mm in the upper Midwest US. Red lines indicate the 0.5 probability contour from the ECMWF ensemble for the day +3 forecasts of > 10 mm (24 h)$^{-1}$.

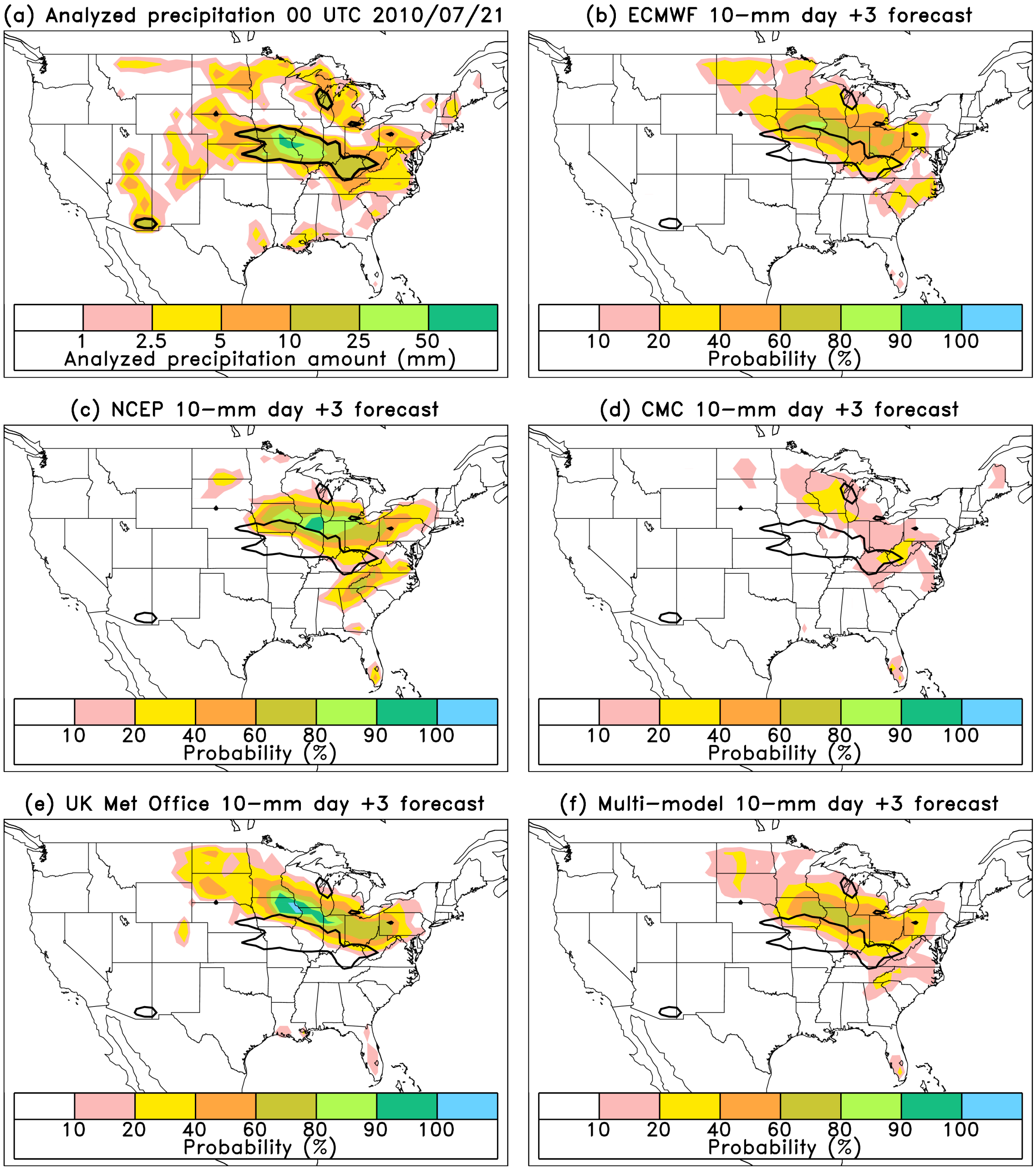
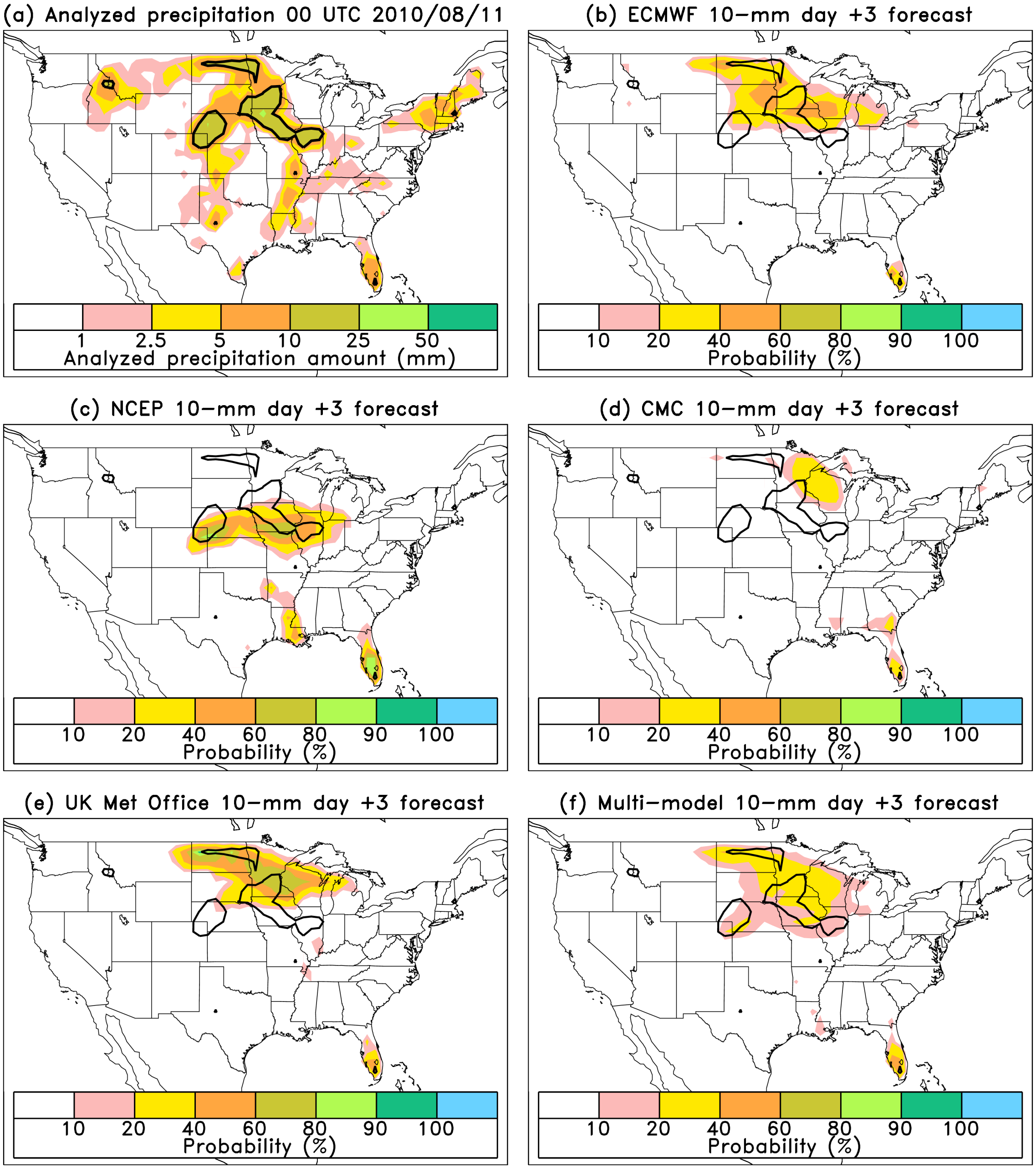

(a) Analyzed precipitation 00 UTC 2010/07/21
1 2.5 5 10 25 50
Analyzed precipitation amount (mm)
(b) ECMWF 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(c) NCEP 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(d) CMC 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(e) UK Met Office 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(f) Multi-model 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)

Figure 7: (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010. 10-mm (24 h)$^{-1}$ contour is denoted by the thick black line. (b) Probability of greater than 10 mm (24 h)$^{-1}$ for day +3 forecast from the ECMWF ensemble for the same period. The analyzed 10-mm contour from panel (a) is repeated. (c) as in (b), but for NCEP. (d) CMC, (e) UK Met Office, and (f) multi-model combination.

1042



(a) Analyzed precipitation 00 UTC 2010/08/11
1 2.5 5 10 25 50
Analyzed precipitation amount (mm)
(b) ECMWF 10–mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(c) NCEP 10–mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(d) CMC 10–mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(e) UK Met Office 10–mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(f) Multi–model 10–mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)

1043
1044
1045 **Figure 8**: As in Fig. 7, but for 24-h period ending 00 UTC 8 August 2010.
1046

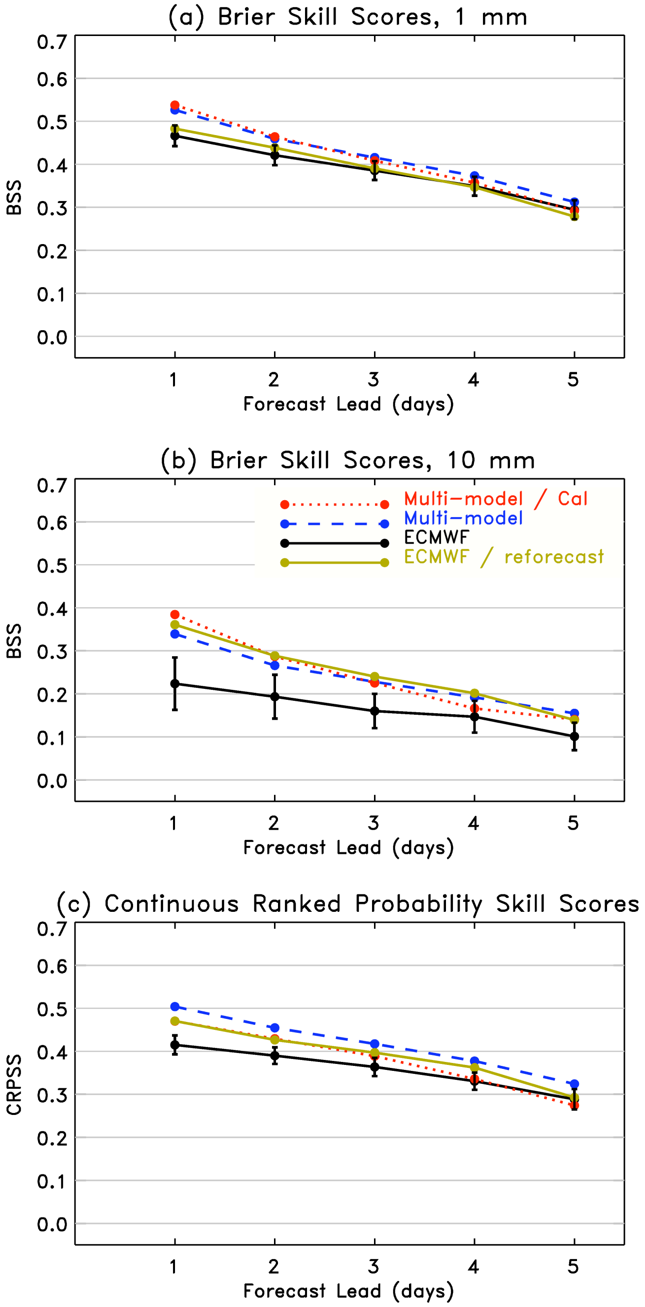
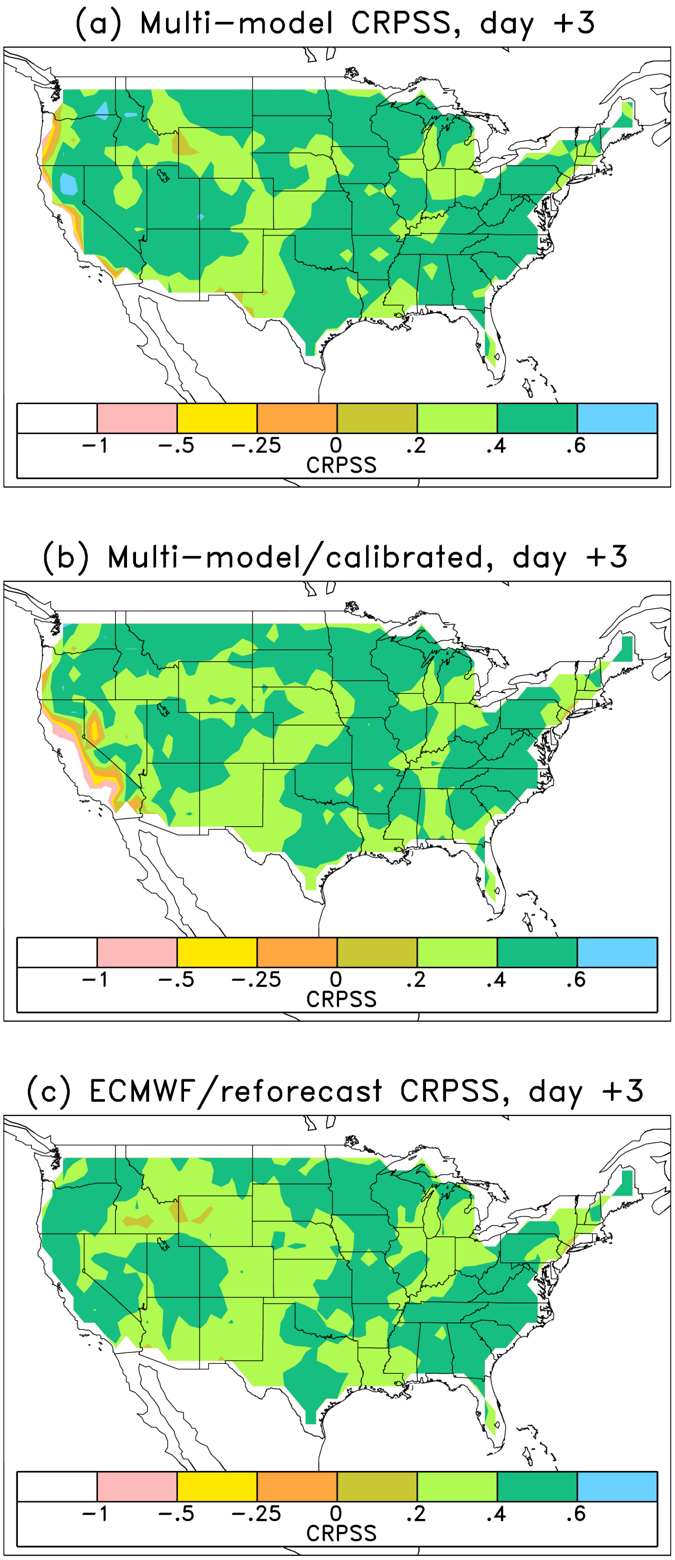


Figure 9: Brier skill scores of various forecasts for (a) > 1-mm (24 h)$^{-1}$ event, and (b) > 10-mm (24 h)$^{-1}$ event, and (c) continuous ranked probability skill scores, all as a function of forecast lead time. "Multi-model/cal" refers to forecasts from the multi-model, calibrated using ELR. "ECMWF/reforecast" refers to ECMWF forecasts calibrated using ELR and the reforecast data set. Error bars denote confidence intervals, the 5th and 95th percentiles of a paired block bootstrap between ECMWF and NCEP forecasts.



(a) Multi-model CRPSS, day +3

(b) Multi-model/calibrated, day +3

(c) ECMWF/reforecast CRPSS, day +3

Figure 10: Maps of *CRPSS* for day +3 forecasts for (a) multi-model, (b) multi-model with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.

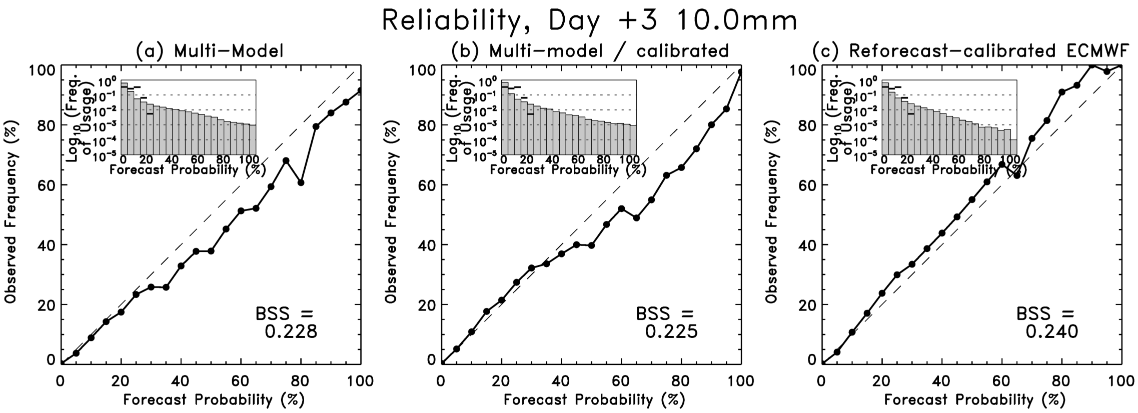


Figure 11: As in Fig. 3, reliability diagrams for day +3 forecasts at the > 10-mm (24 h)$^{-1}$ event. (a) Multi-model forecasts, (b) multi-model with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.

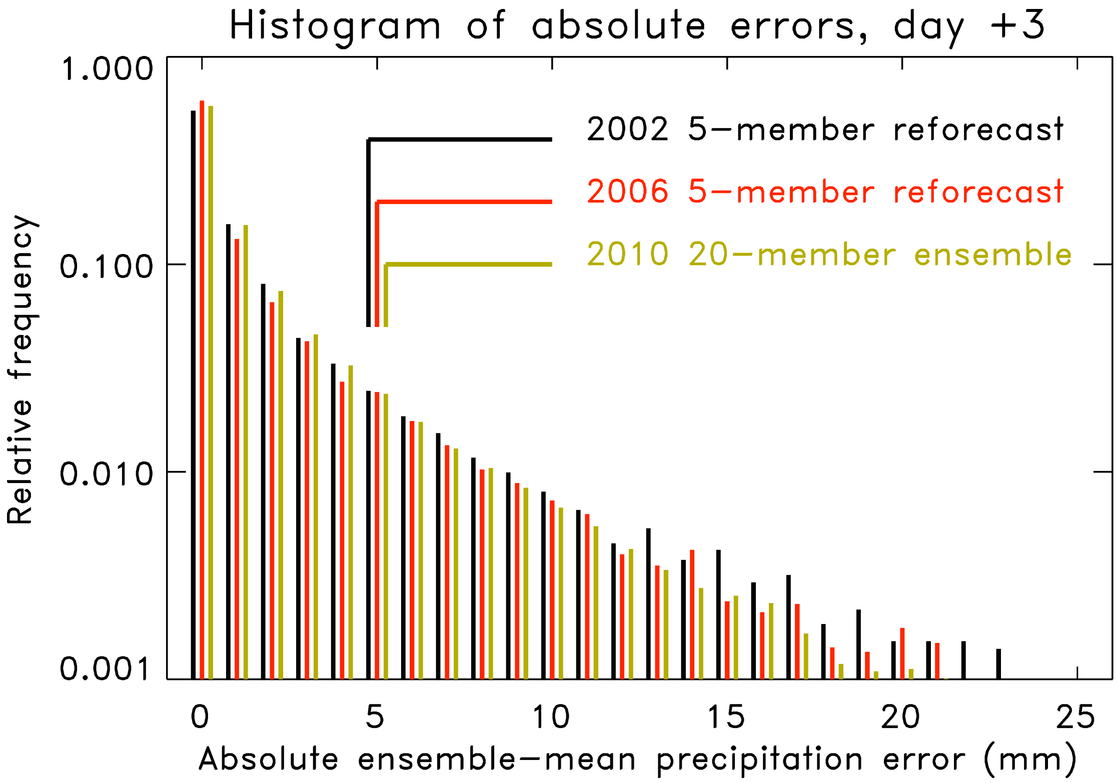


Figure 12: A histogram of the absolute errors of day +3 ensemble-mean precipitation forecasts for the 2002 and 2006 reforecasts and for the 2010, 20-member real-time ensemble.

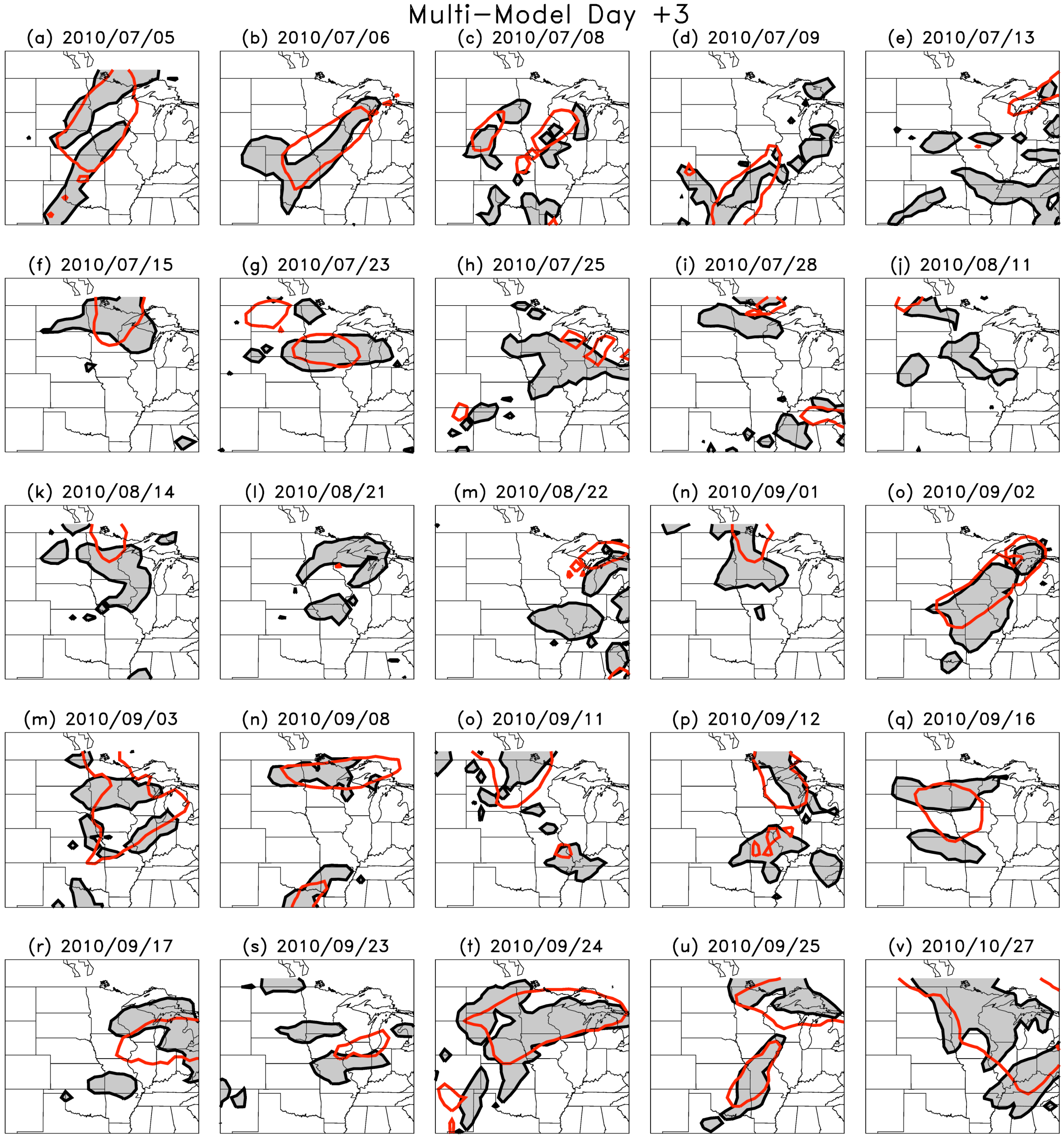


Figure 13: As in Fig. 6, but for multi-model forecasts.

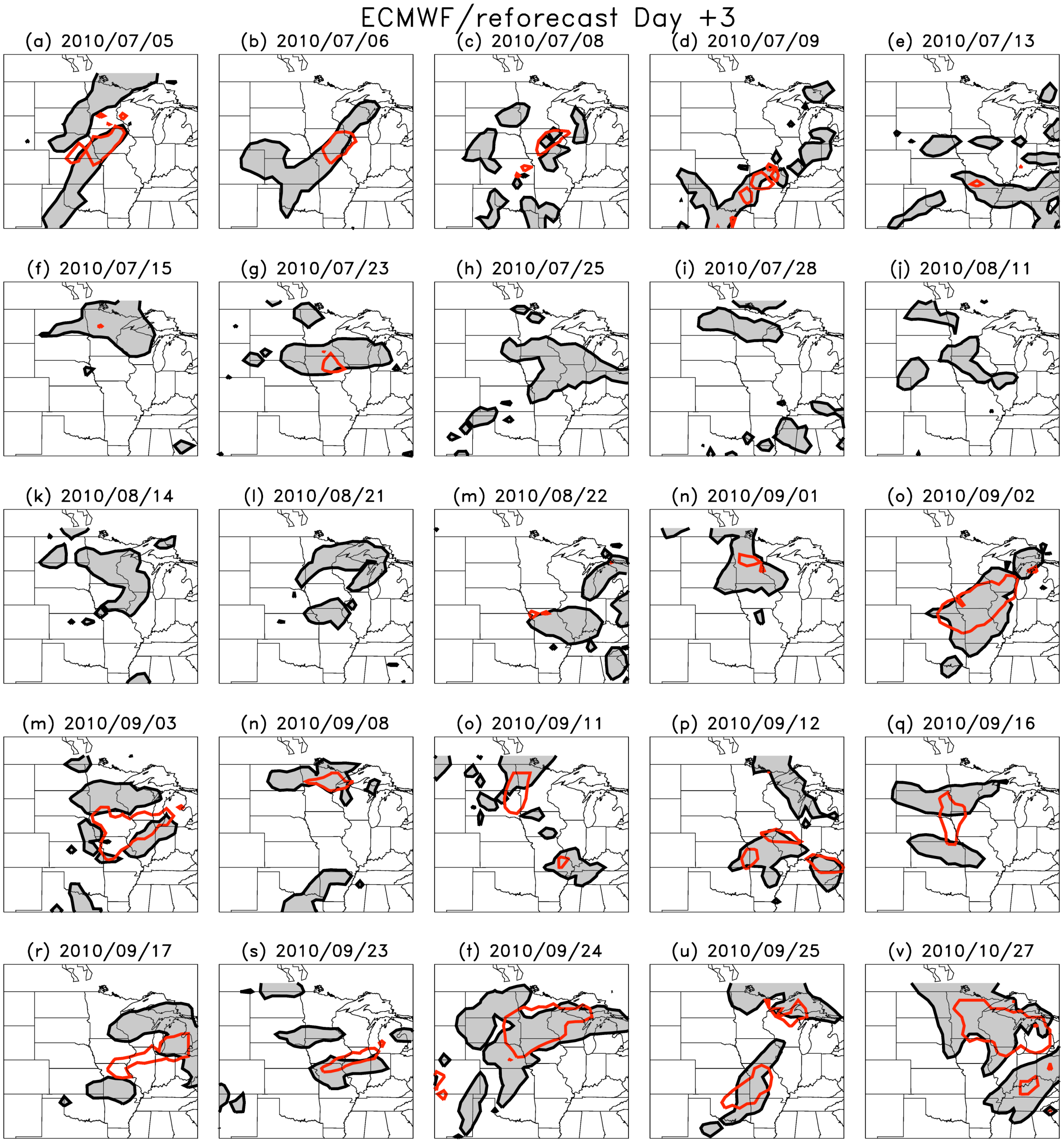


Figure 14: As in Fig. 6, but for reforecast-calibrated ECMWF forecasts.

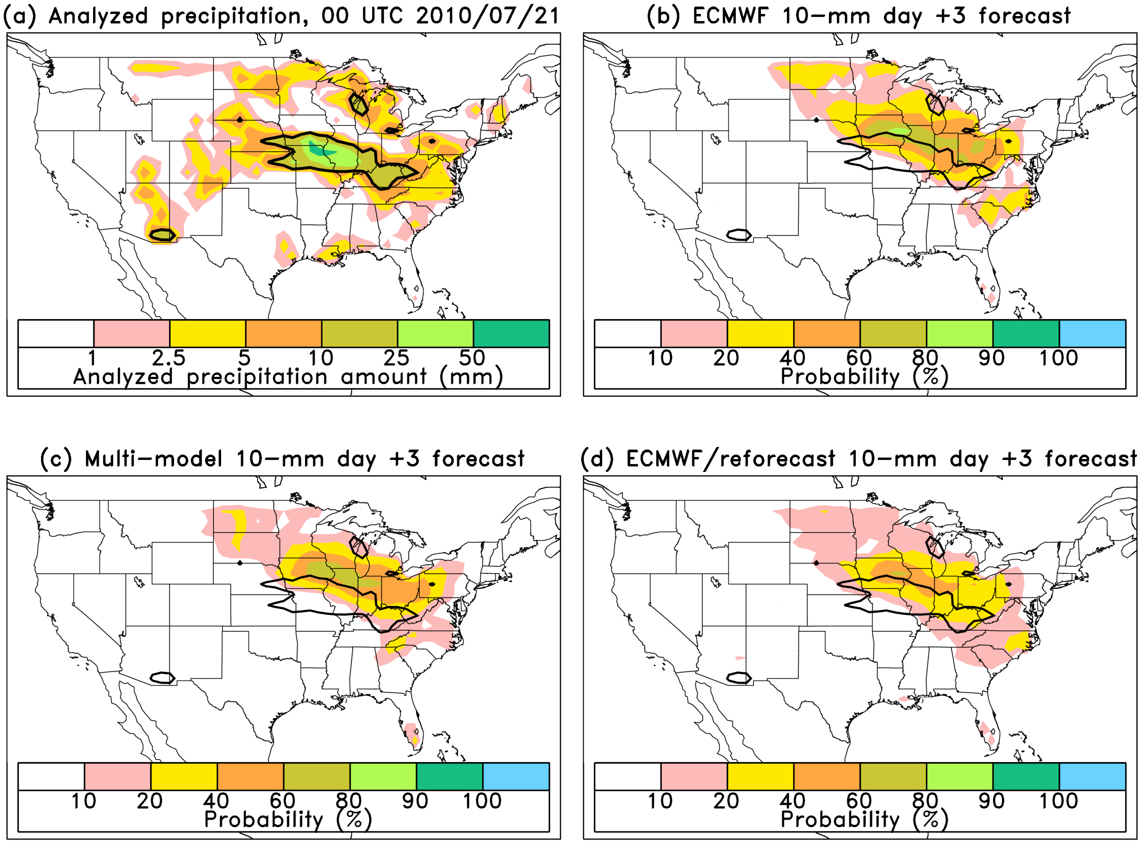

(a) Analyzed precipitation, 00 UTC 2010/07/21
1 2.5 5 10 25 50
Analyzed precipitation amount (mm)
(b) ECMWF 10−mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(c) Multi−model 10−mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(d) ECMWF/reforecast 10−mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)

Figure 15: (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010. 10-mm contour is denoted by the thick black line. (b) Probability of greater than 10 mm $(24\ h)^{-1}$ for day +3 forecast from the ECMWF ensemble for the same period. (c) as in (b), but for multi-model ensemble, and (d) as in (b), but for reforecast-calibrated ECMWF ensemble.

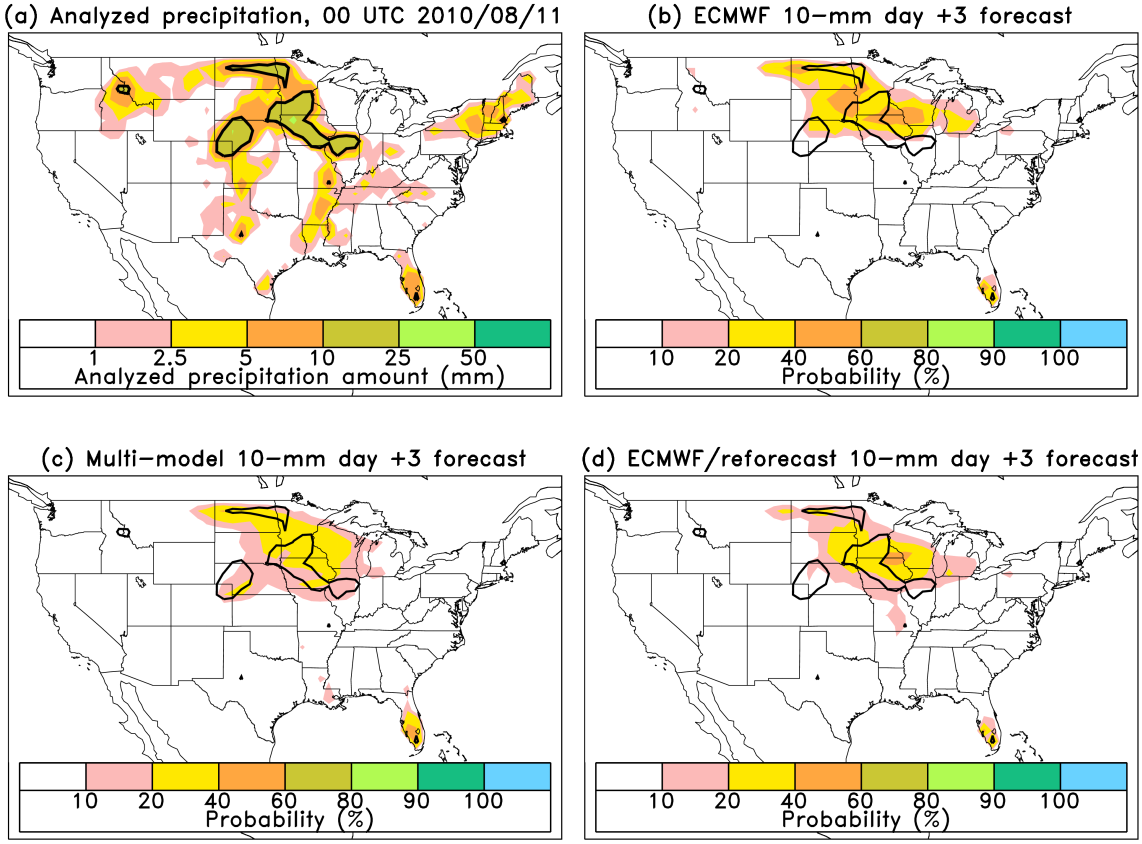

(a) Analyzed precipitation, 00 UTC 2010/08/11
1 2.5 5 10 25 50
Analyzed precipitation amount (mm)
(b) ECMWF 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(c) Multi-model 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)
(d) ECMWF/reforecast 10-mm day +3 forecast
10 20 40 60 80 90 100
Probability (%)

Figure 16: As in Fig. 15, but for the 24-h period ending 00 UTC 8 August 2010.